2008

A kernel-induced space selection approach to model selection of KLDA

Lei Wang
Australian National University, leiw@uow.edu.au

Kap Luk Chan
Nanyang Technological University

Ping Xue
Nanyang Technological University

Luping Zhou
Australian National University, lupingz@uow.edu.au

A kernel-induced space selection approach to model selection of KLDA

Abstract

Model selection in kernel linear discriminant analysis (KLDA) refers to the selection of appropriate parameters of a kernel function and the regularizer. By following the principle of maximum information preservation, this paper formulates the model selection problem as a problem of selecting an optimal kernel-induced space in which different classes are maximally separated from each other. A scatter-matrix-based criterion is developed to measure the "goodness" of a kernel-induced space, and the kernel parameters are tuned by maximizing this criterion. This criterion is computationally efficient and is differentiable with respect to the kernel parameters. Compared with the leave-one-out (LOO) or -fold cross validation (CV), the proposed approach can achieve a faster model selection, especially when the number of training samples is large or when many kernel parameters need to be tuned. To tune the regularization parameter in the KLDA, our criterion is used together with the method proposed by Saadi et al. (2004). Experiments on benchmark data sets verify the effectiveness of this model selection approach.

Keywords

era2015

Disciplines

Engineering | Science and Technology Studies

Publication Details

Wang, L., Chan, K. Luk., Xue, P. & Zhou, L. (2008). A kernel-induced space selection approach to model selection of KLDA. IEEE Transactions on Neural Networks, 19 (12), 2116-2131.

A Kernel-Induced Space Selection Approach to Model Selection in KLDA

Lei Wang, *Member, IEEE*, Kap Luk Chan, *Member, IEEE*, Ping Xue, *Senior Member, IEEE*, and Luping Zhou, *Member, IEEE*

***Abstract*—Model selection in kernel linear discriminant analysis (KLDA) refers to the selection of appropriate parameters of a kernel function and the regularizer. By following the principle of *maximum information preservation*, this paper formulates the model selection problem as a problem of selecting an optimal kernel-induced space in which different classes are maximally separated from each other. A scatter-matrix-based criterion is developed to measure the "goodness" of a kernel-induced space, and the kernel parameters are tuned by maximizing this criterion. This criterion is computationally efficient and is differentiable with respect to the kernel parameters. Compared with the leave-one-out (LOO) or k-fold cross validation (CV), the proposed approach can achieve a faster model selection, especially when the number of training samples is large or when many kernel parameters need to be tuned. To tune the regularization parameter in the KLDA, our criterion is used together with the method proposed by Saadi *et al.* (2004). Experiments on benchmark data sets verify the effectiveness of this model selection approach.**

***Index Terms*—Kernel-induced space selection, kernel linear discriminant analysis (KLDA), kernel parameter tuning, model selection.**

I. Introduction

THE kernel linear discriminant analysis (KLDA or KFDA) incorporates the kernel trick into the linear discriminant analysis (LDA) [2]–[5]. Through a kernel function, data from different classes are implicitly mapped from an input space to a kernel-induced feature space. The LDA is then performed in the kernel-induced feature space to find an optimal direction along which the separability of different classes is maximized. The kernel mapping is often nonlinear, and the dimensionality of the induced feature space can be very high or even infinite [6]. The nonlinearity and the high dimensionality help the KLDA achieve better performance than the LDA, especially when dealing with linearly nonseparable classes. The KLDA has been used in a wide range of practical applications, including feature discovery, data visualization, as well as classification [7]–[9].

Like in other kernel-based learning algorithms, the KLDA also depends on correct model selection. Models that are too complex will overfit training data, whereas oversimplified models cannot effectively capture the underlying structure. Both situations will result in poor classification performance when the KLDA is applied to unseen data. Given a kernel function, model selection for KLDA aims to tune the kernel parameters and the regularization parameter to achieve the best possible discrimination.

Unfortunately, the KLDA cannot do model selection by itself. In other words, the model parameters cannot be tuned by simply maximizing the KLDA's objective function in (2). This is because the KLDA will overfit training data with an unnecessarily complex model, as demonstrated by the experimental study in Fig. 1. When the model parameters are heuristically or empirically set, it is hard to know whether they can lead to sufficiently good discrimination performance. Instead of finding them by trial and error, a systematic and algorithmic approach with sound principles is desired to find the best model parameters.

In the literature, a few criteria have been developed to optimize the model parameters for KLDA [2], [10], [11], [1], [12], [13]. The commonly used k-fold or leave-one-out (LOO) cross-validation (CV) error rate is employed in [2]. The model parameter set that minimizes the error rate is searched. The search technique can be a straightforward exhaustive grid-based search or other more sophisticated ones. Traditionally, to evaluate the k-fold or LOO CV error rate, the KLDA has to be trained and tested on multiple pairs of training and validation subsets. For each model parameter set, the computational complexity of evaluating an LOO CV error rate can reach $\mathcal{O}(n^4)$, where n is the number of training samples.

Efforts have been made to reduce this computational complexity. In [10] and [11], the *Bartlett–Sherman–Woodbury–Morrison* formula is employed to solve a series of matrix inverses in a more efficient way. By doing so, the LOO CV error rate can be evaluated by merely computing the inverse of an $n \times n$ matrix once. This reduces the computational complexity from $\mathcal{O}(n^4)$ to $\mathcal{O}(n^3)$. Along this direction, the work in [11] and [1] uses the Nelder–Mead simplex method to efficiently search for the optimal model parameter set that minimizes the LOO CV error rate. A plus point of the method in [1] is that it allows the regularization parameter to be tuned much more rapidly if the kernel parameters are given. It is worth noting that the LOO CV error rate used in [11] and [1] is differentiable,

Manuscript received March 29, 2007; revised October 18, 2007 and March 25, 2008; accepted May 21, 2008. First published November 17, 2008; current version published November 28, 2008. The early work of this paper was carried out at Nanyang Technological University, Singapore, supported by the Nanyang Technological University under Grants LIT 2002-4 of A-STAR and RGM 14/02. The further work of this paper was carried out at The Australian National University with the support by the Australian Research Council (ARC) Discovery Project under Grant DP0773761.

L. Wang and L. Zhou are with the Research School of Information Sciences and Engineering, The Australian National University, Canberra, A.C.T. 0200, Australia (e-mail: Lei.Wang@rsise.anu.edu.au).

K. Chan and P. Xue are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2008.2005140

and therefore, a gradient-based optimization technique can be applied. In addition, the LOO CV error rate in their work can handle multiple kernel parameter tuning, although the case of single kernel parameter tuning is the focus there. In [12], the LOO CV error rate is modified to be differentiable by approximating a step function with a smooth sigmoid function. In [13], a Bayesian interpretation of the KLDA is provided. The marginal log-likelihood of the data, which is also differentiable, is maximized to tune the model parameters. Nevertheless, each evaluation of the criteria in [11], [1], [12], and [13] still requires the matrix-inverse operation to be performed once, leading to a computational complexity of $\mathcal{O}(n^3)$. When the number of training samples is large or when many kernel parameters are to be tuned, this can still result in a lengthy model selection process, even if the gradient-based optimization technique is used. To make the KLDA applicable to practical applications in which a faster model selection process is desired, there is still room for improvement.

This paper tackles the model selection problem of the KLDA from another perspective. Our key idea is briefly described as follows. It is known that each kernel function corresponds to an implicit mapping from an input space to a kernel-induced feature space. The mapping and the resultant feature space change with the kernel parameter values. Thus, given a kernel function, *tuning the kernel parameters can be interpreted as finding an optimal feature space with which the KLDA can achieve the best discrimination performance*. The optimal feature space can be defined as follows.

In designing an optimal perceptual system, the principle of *maximum information preservation* [14] suggests that such a system should be organized to make the information maximally preserved when passing each processing stage. Recall that the key information in the KLDA is class separability. Applying this principle means that the class separability should be maximally preserved when passing each mapping, including the mapping from an input space to a feature space. This can be intuitively understood because information cannot be recovered in later steps once it is lost. In this sense, an optimal feature space should maximally preserve the separability of classes. Therefore, the kernel parameters can be tuned by maximizing a class separability criterion in a kernel-induced feature space.

In this paper, the commonly used scatter-matrix-based class separability criterion is adopted to measure the class separability in a kernel-induced feature space. This criterion includes the kernel parameters as its functional variables. It has the following properties when being used as a model selection criterion of KLDA. 1) It does not need to perform matrix inverses or to train the KLDA. Once the kernel matrix is ready, this criterion can be quickly evaluated with little computational overhead. 2) It is differentiable as long as the kernel function is, and its derivatives can be easily computed. This makes the criterion quite suitable for the gradient-based optimization technique that is critical for handling a large number of kernel parameters. 3) This criterion is completely rooted in the KLDA. It does not depend on the classifiers or the tasks subsequent to the KLDA. However, this criterion is independent of the regularization parameter, which is also important for the KLDA. As a result, it cannot be used to tune this parameter directly. To circumvent this problem, this criterion is integrated with the method developed in [1]. By doing so, the regularization parameter can be efficiently optimized once the kernel parameters have been tuned. It is worth noting that this scatter-matrix-based criterion was proposed in our previous work in [15] to tune the kernel parameters for support vector machines (SVMs). Such a criterion is also used in [16] to optimize the conformal transformation of a kernel for kernel-based learning algorithms.

The rest of this paper is organized as follows. In Section II, the KLDA is briefly introduced. In Section III, the kernel-induced feature space selection approach is proposed and a class separability criterion is developed in a kernel-induced feature space. The computational complexity and numerical stability of this criterion are discussed. The problem in tuning multiple kernel parameters with this criterion and the relationship between this criterion and the KLDA are also discussed. Section IV presents the experimental study using benchmark data sets. Finally, concluding remarks and future work are given in Section V.

II. Kernel Linear Discriminant Analysis

Let $(\mathbf{x}, y)(\mathbf{x} \in \mathbb{R}^d, y \in \{\pm 1\})$ denote a d-dimensional training sample, where $\mathbb{R}^d$ denotes an input space and y is a class label. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ denote the training sets from the classes of $+1$ and -1, respectively. The sizes of $\mathcal{D}_1$ and $\mathcal{D}_2$ are n_1 and n_2, respectively. Let $\mathcal{D}$ denote the set of all the training samples, and let its size be $n = n_1 + n_2$. In the KLDA, two classes are implicitly mapped from $\mathbb{R}^d$ to a feature space, $\mathcal{F}$. The LDA is performed in $\mathcal{F}$ to find an optimal projection to a subspace $\mathcal{S}$, where the two classes are maximally separated. Let $\phi(\cdot) : \mathbb{R}^d \rightarrow \mathcal{F}$ denote the mapping and $k_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ denote the kernel function, where $\boldsymbol{\theta}$ is the set of kernel parameters and $\langle \cdot, \cdot \rangle$ is the inner product. $\mathbf{K}$ denotes the kernel matrix and $\{\mathbf{K}\}_{i,j}$ is defined as $k_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j)$. Let $\mathbf{K}_{\mathcal{A},\mathcal{B}}$ be a kernel matrix computed with the samples from $\mathcal{A}$ and $\mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ denote two subsets of $\mathcal{D}$. Let $\mathbf{S}_B^\phi$ and $\mathbf{S}_W^\phi$ denote the *between*-class scatter matrix and the *within*-class scatter matrix in $\mathcal{F}$, respectively. They are defined as

$$\mathbf{S}_B = \sum_{i=1}^{2} n_i \left(\mathbf{m}_i^\phi - \mathbf{m}^\phi\right)\left(\mathbf{m}_i^\phi - \mathbf{m}^\phi\right)^\top$$
$$\mathbf{S}_W = \sum_{i=1}^{2} \sum_{\mathbf{x} \in \mathcal{D}_i} \left(\phi(\mathbf{x}) - \mathbf{m}_i^\phi\right)\left(\phi(\mathbf{x}) - \mathbf{m}_i^\phi\right)^\top \quad (1)$$

where $\mathbf{m}_i^\phi$ denotes the mean of the training samples from class i and $\mathbf{m}^\phi$ is the mean of all the training samples in $\mathcal{F}$. The KLDA finds a direction represented by the vector $\mathbf{w}(\mathbf{w} \in \mathcal{F} \setminus \{\mathbf{0}\})$ that maximizes

$$\mathcal{J}(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_B^\phi \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W^\phi \mathbf{w}}. \quad (2)$$

Because the feature space $\mathcal{F}$ is only accessible via the kernel function, this maximization problem cannot be solved by the usual method in the LDA [17]. As pointed out in [2], $\mathbf{w}$ must lie

in the span of all the training samples. Thus, $\mathbf{w}$ is represented as a linear combination of the training samples $\phi(\mathbf{x}_i)(\mathbf{x}_i \in \mathcal{D})$ as

$$\mathbf{w} = \mathbf{\Phi}\boldsymbol{\alpha} \tag{3}$$

where $\mathbf{\Phi}_{\dim(\mathcal{F})\times n}$ is a matrix in which the ith column is $\phi(\mathbf{x}_i)$, and $\boldsymbol{\alpha}_{n\times 1}$ is a vector of expansion coefficients. From this, (2) becomes

$$\mathcal{J}(\mathbf{w}) = \frac{\boldsymbol{\alpha}^\top \mathbf{\Phi}^\top \mathbf{S}_B^\phi \mathbf{\Phi}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \mathbf{\Phi}^\top \mathbf{S}_W^\phi \mathbf{\Phi}\boldsymbol{\alpha}} \triangleq \frac{\boldsymbol{\alpha}^\top \mathbf{P}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \mathbf{Q}\boldsymbol{\alpha}} \tag{4}$$

where $\mathbf{P} \triangleq \mathbf{\Phi}^\top \mathbf{S}_B^\phi \mathbf{\Phi}$ and $\mathbf{Q} \triangleq \mathbf{\Phi}^\top \mathbf{S}_W^\phi \mathbf{\Phi}$. $\mathbf{P}$ and $\mathbf{Q}$ can be fully represented by the kernel function as follows.

$$\mathbf{P} = (\mathbf{M}_1 - \mathbf{M}_2)(\mathbf{M}_1 - \mathbf{M}_2)^\top \tag{5}$$

where $\mathbf{M}_1$ is an n-dimensional vector in which $(\mathbf{M}_1)_j = (1/n_1)\sum_{\mathbf{x}_r \in \mathcal{D}_1} k(\mathbf{x}_j, \mathbf{x}_r)$ and $\mathbf{M}_2$ is obtained in a similar way. The matrix Q is represented as

$$\mathbf{Q} = \sum_{i=1}^{2} \left[\mathbf{K}_{\mathcal{D},\mathcal{D}_i}(\mathbf{I} - \mathbf{1}_{n_i})\mathbf{K}_{\mathcal{D},\mathcal{D}_i}^\top\right] \tag{6}$$

where $\mathbf{I}$ is an identity matrix of size $n_i \times n_i$, and $\mathbf{1}_{n_i}$ is a matrix in which all the elements are $1/n_i$. To ensure numerical stability and to control the learning complexity, a regularized version of $\mathbf{Q}$ is often used as

$$\mathbf{Q}_{\text{reg}} = \mathbf{Q} + \mu\mathbf{I} \tag{7}$$

where μ is the regularization parameter in the KLDA. In this way, $\boldsymbol{\alpha}$ in (4) can be obtained as the eigenvector of $\mathbf{Q}_{\text{reg}}^{-1}\mathbf{P}$ corresponding to the largest eigenvalue. The projection of $\phi(\mathbf{x})$ to the subspace $\mathcal{S}$ is then obtained as

$$\mathbf{x}_p = \mathbf{w}^\top \phi(\mathbf{x}) = \boldsymbol{\alpha}^\top \mathbf{\Phi}^\top \phi(\mathbf{x}) = \boldsymbol{\alpha}^\top \mathbf{K}_{\mathcal{D},\{\mathbf{x}\}} \tag{8}$$

where $\mathbf{x}_p$ denotes the projection. When performing classification, both training and test data are projected to the subspace, and a classifier, such as the Bayes classifier, is applied.

III. The Kernel-Induced Space Selection Approach

The KLDA is a process comprising two projections

$$\mathbf{x} \in \mathbb{R}^d \longrightarrow \phi(\mathbf{x}) \in \mathcal{F} \longrightarrow \mathbf{w}^\top \phi(\mathbf{x}) = \mathbf{x}_p \in \mathcal{S}. \tag{9}$$

A sample $\mathbf{x}$ is successively projected to two spaces $\mathcal{F}$ and $\mathcal{S}$, and finally, becomes $\mathbf{x}_p$. The class separability information is presented at the left end of the process pipeline in (9), and it is hoped that this information could be well preserved when it reaches the right end. This forms a flow of information. Following the principle of *maximum information preservation*, the optimal feature space $\mathcal{F}^*$ should maximally preserve the class separability information. Because the feature space $\mathcal{F}$ changes with the kernel parameter set $\boldsymbol{\theta}$, the model selection for KLDA can be formulated as a feature space selection problem

$$\boldsymbol{\theta}^* \longleftarrow \mathcal{F}_{\boldsymbol{\theta}}^* = \arg\max_{\boldsymbol{\theta}\in\mathbf{\Theta}} [\mathcal{C}(\mathcal{F}_{\boldsymbol{\theta}}, \mathcal{D})] \tag{10}$$

where $\mathbf{\Theta}$ denotes a kernel parameter space and $\mathcal{C}$ is a criterion measuring the class separability in $\mathcal{F}$. The optimal kernel parameter set $\boldsymbol{\theta}^*$ is obtained by maximizing the criterion $\mathcal{C}$.

A. A Realization of This Approach

In the following, the scatter-matrix-based measure is developed in a feature space $\mathcal{F}$ to evaluate the class separability. This measure can take the form of $\text{tr}(\mathbf{S}_B)/\text{tr}(\mathbf{S}_W)$, $|\mathbf{S}_B|/|\mathbf{S}_W|$, or other combinations [17]–[19], where $\text{tr}(\cdot)$ denotes the trace of a matrix and $|\cdot|$ denotes the determinant. Note that the determinant-based measure becomes invalid in this case because the high dimensionality of $\mathcal{F}$ can easily cause these scatter matrices to be singular, resulting in zero determinants. Hence, this paper adopts the trace-based measure and derives the traces of $\mathbf{S}_B^\phi$ and $\mathbf{S}_W^\phi$ below. The superscript ϕ is used to distinguish the variables in $\mathcal{F}$ from those in the input space $\mathbb{R}^d$.

Recall that $\mathbf{m}_1^\phi$ and $\mathbf{m}_2^\phi$ denote the mean vectors of the training samples from the classes of $+1$ and -1 in $\mathcal{F}$. Let $\mathbf{1}$ be a vector whose elements are all "1." Its size will be decided by the context. The following results can be obtained:

$$\begin{aligned}
\mathbf{m}_1^{\phi\top}\mathbf{m}_1^\phi &= n_1^{-2}\cdot \mathbf{1}^\top \mathbf{K}_{\mathcal{D}_1,\mathcal{D}_1}\mathbf{1} \\
\mathbf{m}_2^{\phi\top}\mathbf{m}_2^\phi &= n_2^{-2}\cdot \mathbf{1}^\top \mathbf{K}_{\mathcal{D}_2,\mathcal{D}_2}\mathbf{1} \\
\mathbf{m}_1^{\phi\top}\mathbf{m}_2^\phi &= (n_1 n_2)^{-1}\cdot \mathbf{1}^\top \mathbf{K}_{\mathcal{D}_1,\mathcal{D}_2}\mathbf{1}.
\end{aligned} \tag{11}$$

Based on these, $\text{tr}(\mathbf{S}_B^\phi)$ and $\text{tr}(\mathbf{S}_W^\phi)$ are derived as

$$\begin{aligned}
\text{tr}\left(\mathbf{S}_B^\phi\right) &= \text{tr}\left[\sum_{i=1}^{2} n_i \left(\mathbf{m}_i^\phi - \mathbf{m}^\phi\right)\left(\mathbf{m}_i^\phi - \mathbf{m}^\phi\right)^\top\right] \\
&= \frac{\mathbf{1}^\top \mathbf{K}_{\mathcal{D}_1,\mathcal{D}_1}\mathbf{1}}{n_1} + \frac{\mathbf{1}^\top \mathbf{K}_{\mathcal{D}_2,\mathcal{D}_2}\mathbf{1}}{n_2} - \frac{\mathbf{1}^\top \mathbf{K}_{\mathcal{D},\mathcal{D}}\mathbf{1}}{n}
\end{aligned} \tag{12}$$

$$\begin{aligned}
\text{tr}\left(\mathbf{S}_W^\phi\right) &= \text{tr}\left[\sum_{i=1}^{2}\sum_{\mathbf{x}\in\mathcal{D}_i} \left(\phi(\mathbf{x}) - \mathbf{m}_i^\phi\right)\left(\phi(\mathbf{x}) - \mathbf{m}_i^\phi\right)^\top\right] \\
&= \text{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}}) - \frac{\mathbf{1}^\top \mathbf{K}_{\mathcal{D}_1,\mathcal{D}_1}\mathbf{1}}{n_1} - \frac{\mathbf{1}^\top \mathbf{K}_{\mathcal{D}_2,\mathcal{D}_2}\mathbf{1}}{n_2}.
\end{aligned} \tag{13}$$

The class separability in a feature space $\mathcal{F}$ is obtained as

$$\mathcal{C}(\boldsymbol{\theta}) = \frac{\text{tr}(\mathbf{S}_B^\phi)}{\text{tr}\left(\mathbf{S}_W^\phi\right)}. \tag{14}$$

Thus, the kernel parameter set $\boldsymbol{\theta}^*$ can be optimized as

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}\in\mathbf{\Theta}} \left[\frac{\text{tr}\left(\mathbf{S}_B^\phi\right)}{\text{tr}\left(\mathbf{S}_W^\phi\right)}\right]. \tag{15}$$

Several issues about this criterion are discussed in the following sections.

B. Computational Complexity

The proposed criterion $\mathcal{C}(\boldsymbol{\theta})$ has continuous first- and second-order derivatives with respect to the kernel parameters as long as the kernel function has. Hence, the maximization of $\mathcal{C}(\boldsymbol{\theta})$

can be solved by applying the gradient-based optimization technique. In the optimization process, the computational cost at each iteration is largely due to the evaluation of the criterion $\mathcal{C}(\boldsymbol{\theta})$. This involves the calculation of $\mathrm{tr}(\mathbf{S}_B^\phi)$ and $\mathrm{tr}(\mathbf{S}_W^\phi)$. From (12), it is known that computing $\mathrm{tr}(\mathbf{S}_B^\phi)$ is essentially to calculate $\mathbf{1}^\top\mathbf{K}_{\mathcal{D}_1,\mathcal{D}_1}\mathbf{1}, \mathbf{1}^\top\mathbf{K}_{\mathcal{D}_2,\mathcal{D}_2}\mathbf{1}$, and $\mathbf{1}^\top\mathbf{K}_{\mathcal{D}_1,\mathcal{D}_2}\mathbf{1}$. Although written in the matrix form, they are simply the summation of all the entries in each of the kernel matrices $\mathbf{K}_{\mathcal{D}_1,\mathcal{D}_1}, \mathbf{K}_{\mathcal{D}_2,\mathcal{D}_2}$, and $\mathbf{K}_{\mathcal{D}_1,\mathcal{D}_2}$. Computing them requires n_1^2, n_2^2, and $n_1 n_2$ additions, respectively. Similarly, calculating the three terms of $\mathrm{tr}(\mathbf{S}_W^\phi)$ in (13) requires n, n_1^2, and n_2^2 additions, respectively. Once $\mathrm{tr}(\mathbf{S}_B^\phi)$ and $\mathrm{tr}(\mathbf{S}_W^\phi)$ are obtained, the criterion $\mathcal{C}(\boldsymbol{\theta})$ can be instantly computed by a single division. Therefore, the computational complexity of $\mathcal{C}(\boldsymbol{\theta})$ is no more than $\mathcal{O}(n^2)$ with the basic operation of *addition*. It is much less than the computational complexity of the criteria that need to compute a matrix inverse or to train a KLDA [1], [11]–[13]. They result in a complexity of $\mathcal{O}(n^3)$ with the basic operation of *multiplication*. Hence, it can be expected that model selection with the proposed criterion $\mathcal{C}(\boldsymbol{\theta})$ is faster. This will be verified by the experimental study later. Meanwhile, we would like to point out that the total computational cost in model selection is also affected by the number of iterations and the number of function evaluations in the optimization process, as well as the dimensionality of the input space and the complexity of the kernel function, which are regarded as constants for a given problem.

C. Relationship to the Goal of KLDA

From the definition of $\mathbf{S}_B$ and $\mathbf{S}_W$ in (1), we know that they are positive semidefinite (PSD). Following the property of Rayleigh quotient, it can be obtained that

$$\begin{aligned} 0 \le \frac{\mathbf{w}^\top\mathbf{S}_B^\phi\mathbf{w}}{\mathbf{w}^\top\mathbf{w}} \le \lambda_{\max}\left(\mathbf{S}_B^\phi\right) \\ 0 \le \frac{\mathbf{w}^\top\mathbf{S}_W^\phi\mathbf{w}}{\mathbf{w}^\top\mathbf{w}} \le \lambda_{\max}\left(\mathbf{S}_W^\phi\right) \end{aligned} \tag{16}$$

where $\lambda_{\max}(\mathbf{S}_B^\phi)$ and $\lambda_{\max}(\mathbf{S}_W^\phi)$ denote the maximum eigenvalues of $\mathbf{S}_B^\phi$ and $\mathbf{S}_W^\phi$, respectively. Following this, the objective function of the KLDA can be expressed as

$$\begin{aligned} \mathcal{J}(\mathbf{w}) = \frac{\mathbf{w}^\top\mathbf{S}_B^\phi\mathbf{w}}{\mathbf{w}^\top\mathbf{S}_W^\phi\mathbf{w}} = \frac{\mathbf{w}^\top\mathbf{S}_B^\phi\mathbf{w}/\mathbf{w}^\top\mathbf{w}}{\mathbf{w}^\top\mathbf{S}_W^\phi\mathbf{w}/\mathbf{w}^\top\mathbf{w}} \\ \ge \frac{\mathbf{w}^\top\mathbf{S}_B^\phi\mathbf{w}/\mathbf{w}^\top\mathbf{w}}{\lambda_{\max}\left(\mathbf{S}_W^\phi\right)}. \end{aligned} \tag{17}$$

Hence

$$\begin{aligned} &\max_{\mathbf{w}\in\mathcal{F}\setminus\{\mathbf{0}\}}\left[\mathcal{J}(\mathbf{w})\right] \\ &\ge \max_{\mathbf{w}\in\mathcal{F}\setminus\{\mathbf{0}\}}\left(\frac{\mathbf{w}^\top\mathbf{S}_B^\phi\mathbf{w}/\mathbf{w}^\top\mathbf{w}}{\lambda_{\max}\left(\mathbf{S}_W^\phi\right)}\right) \\ &= \frac{\lambda_{\max}\left(\mathbf{S}_B^\phi\right)}{\lambda_{\max}\left(\mathbf{S}_W^\phi\right)} \ge \frac{\mathrm{tr}(\mathbf{S}_B^\phi)}{\mathrm{tr}\left(\mathbf{S}_W^\phi\right)} = \mathcal{C}(\boldsymbol{\theta}). \end{aligned} \tag{18}$$

The last inequality is based on the following two facts. 1) In a binary classification, $\mathrm{rank}(\mathbf{S}_B^\phi) = 1$ and $\mathbf{S}_B^\phi$ has one and only one nonzero eigenvalue. Thus, it can be obtained that $\lambda_{\max}(\mathbf{S}_B^\phi) = \mathrm{tr}(\mathbf{S}_B^\phi)$. 2) It is known that $\sum_{i=1}^{\dim(\mathcal{F})}\lambda_i(\mathbf{S}_W^\phi) = \mathrm{tr}(\mathbf{S}_W^\phi)$ and that $\lambda_i(\mathbf{S}_W^\phi) \ge 0$ because $\mathbf{S}_W^\phi$ is PSD. Therefore, it can be shown that $0 \le \lambda_{\max}(\mathbf{S}_W^\phi) \le \mathrm{tr}(\mathbf{S}_W^\phi)$. The result of (18) indicates that the criterion $\mathcal{C}$ in (15) is essentially a lower bound of the maximum value of KLDA's objective function. Because the goal of the KLDA is to maximize its objective function, maximizing the proposed criterion $\mathcal{C}$ for model selection is consistent with this goal.

The relationship between the KLDA and the criterion $\mathcal{C}$ is summarized as follows. 1) Because the KLDA cannot perform model selection automatically, the criterion $\mathcal{C}$ is proposed to accomplish this task. In other words, this criterion serves the goal of KLDA. 2) Both KLDA and $\mathcal{C}$ seek the maximization of the class separability. However, the criterion $\mathcal{C}$ finds an optimal higher dimensional feature space $\mathcal{F}^\star$, whereas the KLDA seeks an optimal 1-D subspace $\mathcal{S}^\star$. Their goals are different. 3) The criterion $\mathcal{C}$ does not conflict with the KLDA. It is only used to perform model selection and it cannot replace the KLDA. From the above analysis, it can be said that the proposed criterion $\mathcal{C}$ is *not* an reinvention of the KLDA. Asides from these, this criterion was related to the radius-margin bound of SVMs [20] in our previous work in [15]. Comparatively, its relationship to the KLDA is more essential.

D. Numerical Stability

A good and reliable criterion has to ensure numerical stability when its variables go to extreme values. For example, when a Gaussian radial basis function (GRBF) kernel[1] is used, both $\mathrm{tr}(\mathbf{S}_B^\phi)$ and $\mathrm{tr}(\mathbf{S}_W^\phi)$ will approach zero with the increasing value of the Gaussian width σ. Geometrically, this means that all the training samples are being projected to a single point in $\mathcal{F}$. In this case, the value of the criterion $\mathcal{C}$ becomes indeterminate. In the following, two approaches are developed to ensure numerical stability.

The first method realizes this by deriving a lower bound of $\mathcal{C}$ and using this bound for model selection. Maximizing $(\mathrm{tr}(\mathbf{S}_B^\phi))/(\mathrm{tr}(\mathbf{S}_W^\phi))$ is equivalent to maximizing $(\mathrm{tr}(\mathbf{S}_B^\phi))/(\mathrm{tr}(\mathbf{S}_T^\phi))$, where $\mathbf{S}_T^\phi$ is the total scatter matrix and $\mathrm{tr}(\mathbf{S}_T^\phi) = \mathrm{tr}(\mathbf{S}_B^\phi) + \mathrm{tr}(\mathbf{S}_W^\phi)$. Let k_s denote a stationary kernel. It is defined as a kernel whose value only depends on the difference of the two inputs, that is, $k_s(\mathbf{x}_i, \mathbf{x}_j) = k_s(\mathbf{x}_i - \mathbf{x}_j)$ [21]. Furthermore, let us consider the stationary kernel satisfying $k_s(\mathbf{x}_i, \mathbf{x}_j) \ge 0, \forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}$. The GRBF kernel is just an example of such a stationary kernel. Geometrically, via a stationary kernel, all the training samples are mapped onto a hypersphere in $\mathcal{F}$ with the radius of $\sqrt{k_s(0)}$. This is because $\|\phi(\mathbf{x})\|^2 = k_s(\mathbf{x}, \mathbf{x}) = k_s(\mathbf{x} - \mathbf{x}) = k_s(0)$. In this case, it can be shown that

$$\begin{aligned} \mathrm{tr}(\mathbf{S}_T^\phi) &= \mathrm{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}}) - \frac{\mathbf{1}^\top\mathbf{K}_{\mathcal{D},\mathcal{D}}\mathbf{1}}{n} \\ &\le \mathrm{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}}) - \frac{\mathrm{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{n} = (n-1)k_s(0) \end{aligned} \tag{19}$$

[1]A GRBF kernel is defined as $k(\mathbf{x}_i, \mathbf{x}_j) = \exp((-\|\mathbf{x}_i - \mathbf{x}_j\|^2)/(2\sigma^2))$, where σ is the Gaussian width.

where $k_s(0)$ is constant for a stationary kernel. When the criterion $\mathcal{C}$ is represented as $\mathcal{C}(\boldsymbol{\theta}) = (\mathrm{tr}(\mathbf{S}_B^\phi))/(\mathrm{tr}(\mathbf{S}_T^\phi))$, a lower bound of $\mathcal{C}(\boldsymbol{\theta})$ can be obtained as

$$\mathcal{C}_l(\boldsymbol{\theta}) = \frac{\mathrm{tr}(\mathbf{S}_B^\phi)}{(n-1)k_s(0)} \leq \mathcal{C}(\boldsymbol{\theta}). \tag{20}$$

This suggests that when a stationary kernel is used, maximizing $\mathcal{C}$ can be approximately achieved by maximizing $\mathrm{tr}(\mathbf{S}_B^\phi)$. This avoids the problem of numerical instability in using the quotient of $(\mathrm{tr}(\mathbf{S}_B^\phi))/(\mathrm{tr}(\mathbf{S}_W^\phi))$.

The second method is by appending an extra term to $\mathrm{tr}(\mathbf{S}_W^\phi)$ to ensure numerical stability. Following the two-norm soft margin in SVMs [22], this is conveniently realized by slightly modifying the kernel function as

$$\widetilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} k(\mathbf{x}_i, \mathbf{x}_j), & \text{when } i \neq j \\ k(\mathbf{x}_i, \mathbf{x}_j) + \mu', & \text{when } i = j \end{cases} \tag{21}$$

where μ' is a small positive real number. The resultant criterion is called $\mathcal{C}_{\mu'}$ in this paper. Compared with $\mathcal{C}_l$ in the first method, $\mathcal{C}_{\mu'}$ is closer to the original criterion $\mathcal{C}$ because it does not involve any approximation. The price to pay is an extra parameter μ'. $\mathcal{C}_l$ and $\mathcal{C}_{\mu'}$ form two variants of the proposed criterion for model selection. Both of them will be investigated in the experimental study.

E. Tuning Multiple Kernel Parameters

In general, a kernel function with multiple parameters can be a rather complex learning model. To prevent it from overfitting training samples, regularization is often needed when tuning multiple kernel parameters. Please note that the regularization is generally required in optimizing a criterion with multiple free parameters, although it may be realized in various ways. Without a proper regularization, the optimal solution may overfit the noise in training samples, especially when the number of training samples is small [23]. This situation has been observed in [12] where the LOO error rate is used as a criterion. In this paper, a regularized $\mathcal{C}$ is developed as follows to address this problem:

$$\mathcal{C}_{\mathrm{reg}}(\boldsymbol{\theta}) = (1-\lambda)\mathcal{C}(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \tag{22}$$

where $\lambda (0 \leq \lambda < 1)$ is the regularization parameter that penalizes the deviation of $\boldsymbol{\theta}$ from a preset $\boldsymbol{\theta}_0$. Mathematically, this imposes a Gaussian prior over the parameter set $\boldsymbol{\theta}$, and the mean of this Gaussian distribution is $\boldsymbol{\theta}_0$. When there is no *a priori* knowledge about $\boldsymbol{\theta}$, setting $\boldsymbol{\theta}_0 = \mathbf{0}$ seems to be a good option. In model selection for KLDA, a better setting of $\boldsymbol{\theta}_0$ can be obtained as follows. For the kernels where each feature component is assigned a kernel parameter, for example, the ellipsoidal GRBF kernel,[2] $\boldsymbol{\theta}_0$ can be chosen by imposing the constraint of $\theta_1 = \cdots = \theta_d$ and solving the following optimization:

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{C}(\boldsymbol{\theta})|_{\theta_1=\theta_2=\cdots=\theta_d}. \tag{23}$$

[2]An ellipsoidal GRBF kernel is defined as $k(\mathbf{x}, \mathbf{y}) = \exp(-\sum_{i=1}^{d}((x_i - y_i)^2)/(2\sigma_i^2))$, where σ_i is for the ith feature component.

This constraint reduces the number of free parameters from d to one, and therefore, minimizing $\mathcal{C}$ in this case is less likely to suffer from overfitting. We believe that using $\boldsymbol{\theta}_0$ obtained in this way will be more sensible than simply setting $\boldsymbol{\theta}_0 = \mathbf{0}$. For instance, with the constraint of $\theta_1 = \cdots = \theta_d$, the ellipsoidal GRBF kernel reduces to a common spherical GRBF kernel that only has one kernel parameter. Essentially, $\boldsymbol{\theta}_0$ obtained in (23) has been a well-tuned kernel parameter for the spherical GRBF kernel, subject to the criterion $\mathcal{C}$. By straightforwardly setting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, sufficiently good discriminant performance can be achieved, whereas setting $\boldsymbol{\theta} = \mathbf{0}$ will not. In other words, the former setting secures a good initial value for the kernel parameters. When tuning multiple kernel parameters, $\theta_1, \cdots, \theta_d$ are then allowed to moved around $\boldsymbol{\theta}_0$ to minimize the criterion $\mathcal{C}$ further. The regularization parameter λ needs to be set before tuning multiple kernel parameters. Empirically, the larger the number of free parameters or the smaller the number of training samples is, the larger the λ value should be. This is because overfitting is more likely to occur in these situations. This paper follows this empirical rule to set the λ.

F. Tuning the Regularization Parameter in the KLDA

Before ending Section III, we would like to mention that the proposed criterion $\mathcal{C}$ cannot be used to tune the regularization parameter μ in the KLDA [defined in (7)]. As seen from (11)–(14), the criterion $\mathcal{C}$ is not a function of μ. This is because $\mathcal{C}$ works in the kernel-induced feature space and it does not involve estimating the optimal projection $\mathbf{w}$ in which μ has to be preset. To deal with this problem, this paper integrates the proposed criterion with the method developed in [1]. When the kernel parameters are given, the method in [1] can be used to evaluate the LOO CV error rate for a given μ with a computational complexity of $\mathcal{O}(n^3)$. This allows it to quickly tune this regularization parameter. Model selection for KLDA in our work has two steps. First, the proposed criterion $\mathcal{C}$ is maximized to tune the kernel parameters. After that, these kernel parameters are fed to the method in [1]. Based on these kernel parameters, the LOO CV error rate is minimized to find the optimal regularization parameter μ. Because the method of [1] is not the contribution of this paper, it will not be elaborated here and the readers are referred to the original paper.

IV. Experimental Results

The experiments aim to evaluate the effectiveness of the proposed criterion for model selection in KLDA. Thirteen benchmark data sets in [2], [11], and [1] are used here. They are listed in Table I, where d denotes the dimensionality of an input space and n_{train} and n_{test} are the sizes of training and test sets, respectively. Each data set has been randomly split into 100 pairs of training and test subsets (about 60% : 40%). Note that there are only 20 pairs for the data sets of "Image" and "Splice." Two forms of the GRBF kernel are used. The first form is $k(\mathbf{x}, \mathbf{y}) = \exp(-(\|\mathbf{x} - \mathbf{y}\|^2)/(2\sigma^2))$, where σ is the kernel parameter known as the Gaussian width. In this form, a single σ is uniformly applied to all the feature components, and therefore, this kernel is often called the *spherical* GRBF kernel. In this case, the kernel parameter set is merely $\boldsymbol{\theta} = \{\sigma\}$.

TABLE I
ATTRIBUTES OF THE 13 BENCHMARK DATA SETS

	Bana.	B.Can	Diab.	F.Sol	Germ.	Hear.	Imag.	Spli.	Thyr.	Tita.	Ring.	Twon.	Wave.
d	2	9	8	9	20	13	18	60	5	3	20	20	21
n_{train}	400	200	468	666	700	170	1300	1000	140	150	400	400	400
n_{test}	4900	77	300	400	300	100	1010	2175	75	2051	7000	7000	4600

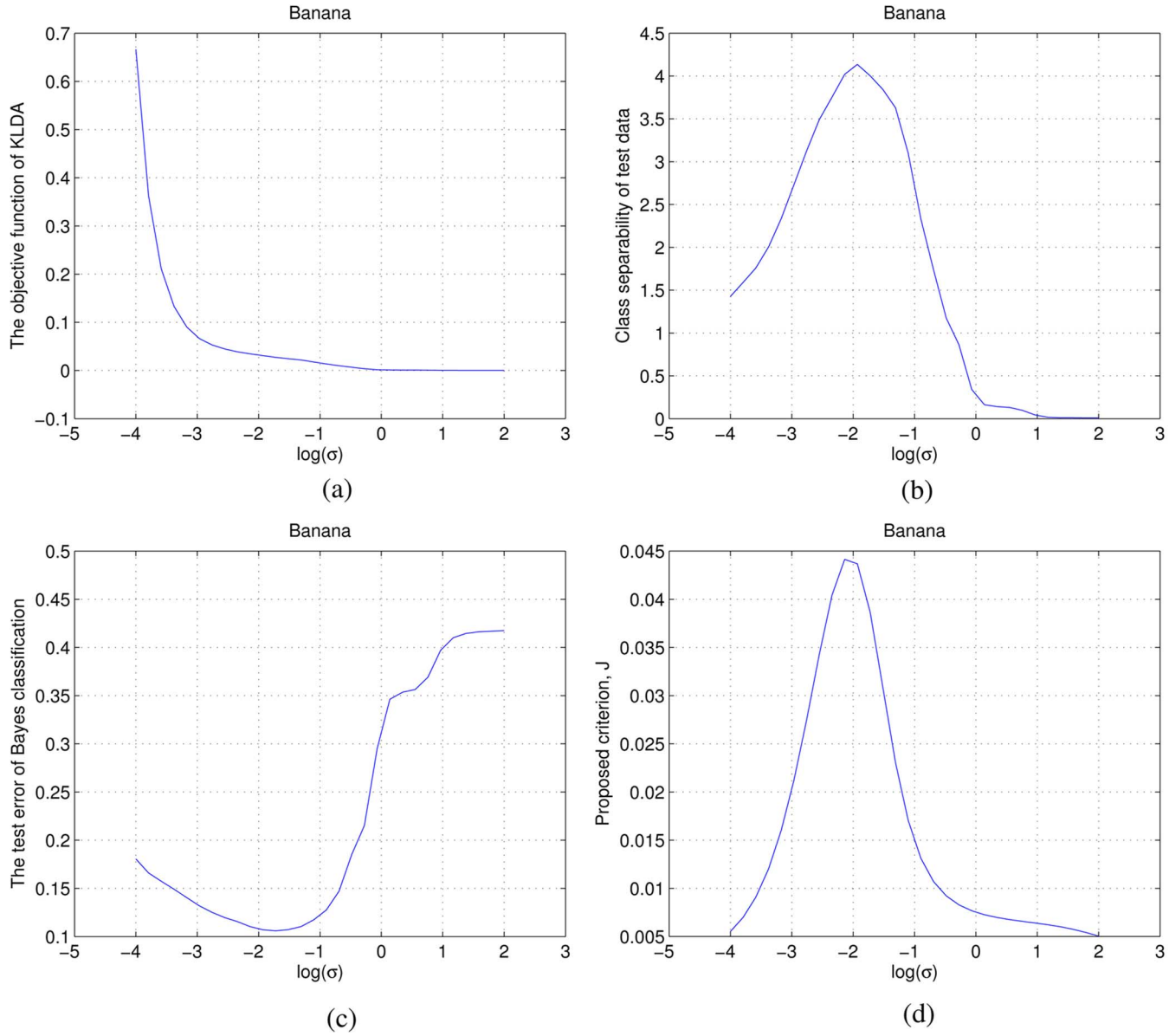


Fig. 1. Demonstration of the properties of the proposed criterion on the *Banana* data set. This figure shows that the KLDA's objective function cannot be used to tune the kernel parameter σ. In contrast, the proposed criterion gives a well-tuned σ ($\approx \exp(-2)$), with which a high class separability is achieved on the test data and a low classification error rate is obtained accordingly. (a) The value of KLDA's object function. (b) Class separability of test data after KLDA. (c) Classification error on test data after KLDA. (d) The value of the criterion $\mathcal{C}_{\mu'}$.

The second form is the *ellipsoidal* GRBF kernel, which assigns different σ values to the d feature components. It is expressed as $k(\mathbf{x},\mathbf{y}) = \exp(-\sum_{i=1}^{d}((x_i - y_i)^2)/(2\sigma_i^2)$, where σ_i is for the ith feature component. The kernel parameter set now expands to $\{\sigma_1, \sigma_2, \cdots, \sigma_d\}$. These two kernels are used to investigate the performance of our criterion in handling single and multiple kernel parameters. Training and testing of the KLDA are done by using the codes written in Matlab. Two variants of the proposed criterion, $\mathcal{C}_{\mu'}$ and $\mathcal{C}_l$, are investigated. μ' in $\mathcal{C}_{\mu'}$ is empirically set to 10^{-3} (may be suboptimal). The parameter λ in (22) for tuning multiple kernel parameters is empirically selected from the range $[0.5, 0.99]$. The regularization parameter μ in the KLDA will be optimally tuned by incorporating the method in [1]. To simplify the optimization in model selection, η_i is used to denote $1/(2\sigma_i^2)$ in the GRBF kernel, and optimizing σ_i becomes the optimization of η_i. Because η_i must be positive, $\log(\eta_i)$ is optimized instead to avoid solving a constrained optimization problem. The initial value of η_i is com-

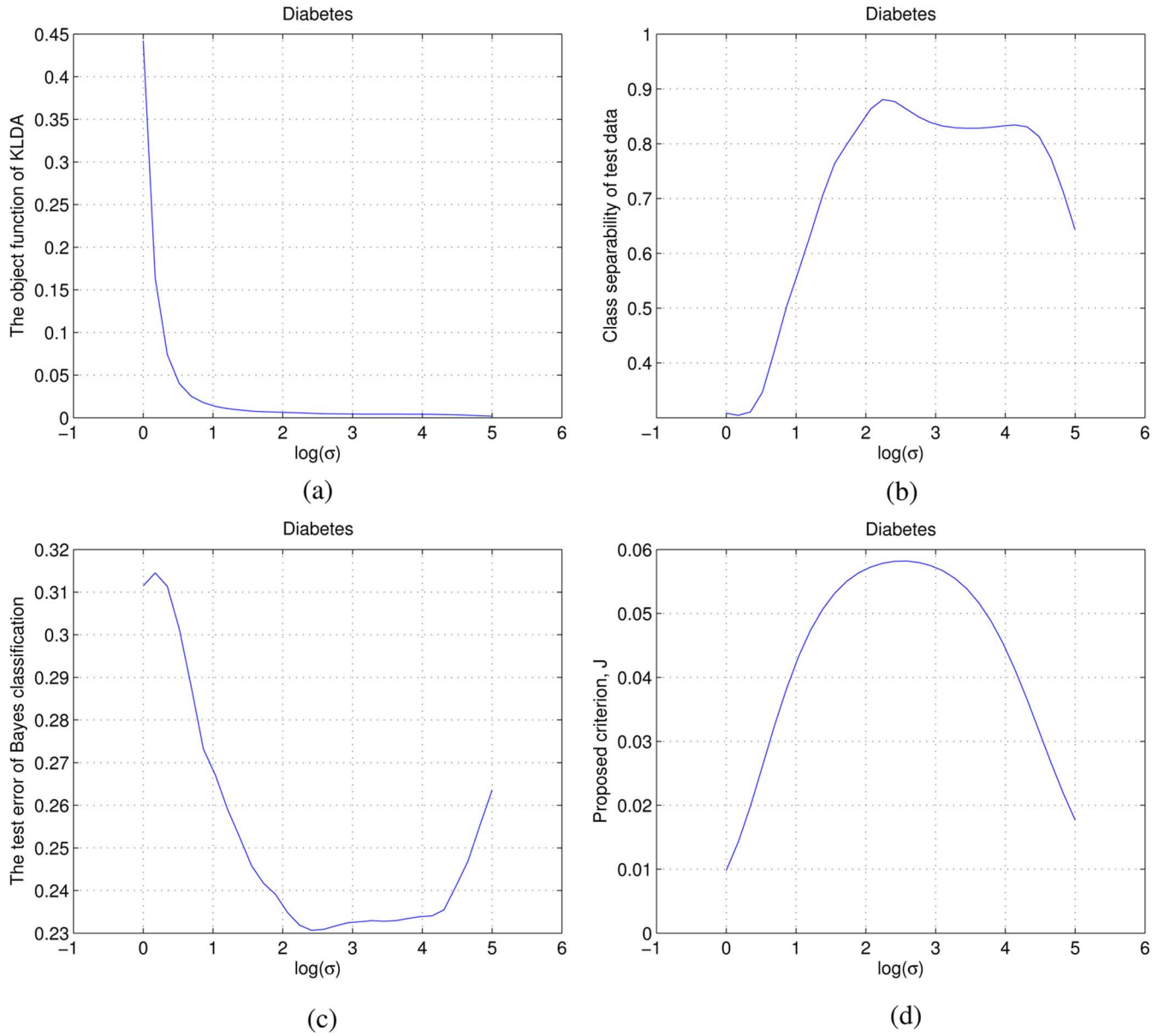


Fig. 2. Demonstration of the properties of the proposed criterion on the *Diabetes* data set. (a) The value of KLDA's object function. (b) Class separability of test data after KLDA. (c) Classification error on test data after KLDA. (d) The value of the criterion $\mathcal{C}_{\mu'}$.

puted by setting $\sigma_i = 0.25 * \sqrt{d}$, where d is the dimensionality of an input space. Note that each feature component of the training data has been linearly scaled to $[0, 1]$ before performing model selection. The test data will be scaled accordingly when performing classification. Both criteria $\mathcal{C}_{\mu'}$ and $\mathcal{C}_l$ are compared with the method proposed in [11], a state-of-the-art model selection technique for the KLDA. In that method, the LOO error rate is efficiently evaluated with a computational complexity of $\mathcal{O}(n^3)$. It has demonstrated excellent performance in tuning the Gaussian width σ and the regularization parameter μ. This experiment will check whether our criteria can give rise to a faster model selection than the method in [11]. The experiments consist of three parts: 1) demonstration of the properties of the proposed criterion, 2) comparison of model selection time and classification error rate for tuning a *single* kernel parameter, and 3) comparison of model selection time and classification error rate for tuning *multiple* kernel parameters.

A. Demonstration of the Properties of the Proposed Criterion

At first, on the data set of "Banana," the KLDA's objective function, the class separability of test data in the 1-D subspace $\mathcal{S}$, the classification error rate in $\mathcal{S}$, and the criterion $\mathcal{C}_{\mu'}$ are plotted against σ in the spherical GRBF kernel. As shown in Fig. 1, all the horizontal axes are in the natural logarithm of σ. Fig. 1(a) shows the value of KLDA's objective function, which indicates the class separability of the *training* data in $\mathcal{S}$. It can be seen that its value monotonically increases with the decreasing value of σ, rather than showing a clear peak. For the spherical GRBF kernel, a smaller σ value often means a more complex mapping function. By comparing this result with the class separability of the *test* data in Fig. 1(b), the effect of "overfitting" can be clearly seen. That is, with a smaller σ [for example, $\log(\sigma) < -2$], the KLDA's objective function value goes up, indicating that a larger class separability has been achieved on

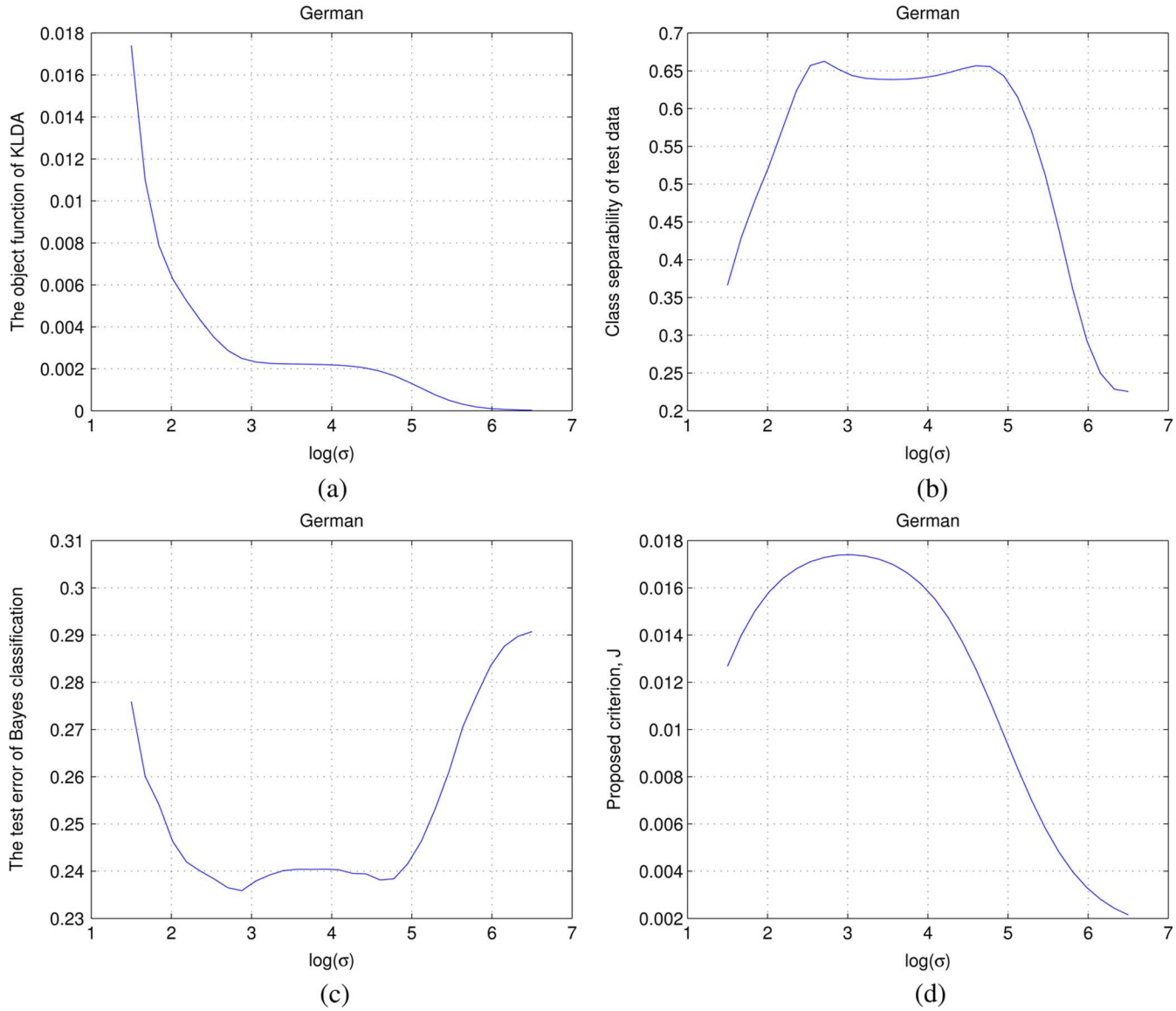


Fig. 3. Demonstration of the properties of the proposed criterion on the *German* data set. (a) The value of KLDA's object function. (b) Class separability of test data after KLDA. (c) Classification error on test data after KLDA. (d) The value of the criterion $\mathcal{C}_{\mu'}$.

the *training* data in $\mathcal{S}$. However, the class separability of the *test* data in $\mathcal{S}$ quickly falls at this time, and the classification error rate in Fig. 1(c) becomes higher. This indicates that the KLDA's objective function cannot be used to tune σ. Fig. 1(d) shows the value of the criterion $\mathcal{C}_{\mu'}$. As seen, its maximum aligns well with the maximum of the class separability of test data [plotted in Fig. 1(b)] and the minimum of classification error rate [plotted in Fig. 1(c)]. This suggests that maximizing the criterion $\mathcal{C}_{\mu'}$ can give a well-tuned σ. Similar results from the data sets of "Diabetes" and "German" are shown in Figs. 2 and 3.

Before starting model selection, the time taken by a single evaluation of $\mathcal{C}_{\mu'}$ or the LOO error rate in [11] is compared in Fig. 4. The horizontal axis is the order of the data sets listed in Table I, while the vertical axis is the evaluation time. As seen, each evaluation of the criterion $\mathcal{C}_{\mu'}$ costs less time than an evaluation of the LOO error rate, showing its advantage of computational efficiency. The above results demonstrate the properties of our criterion and its effectiveness.

The rest of the experiments give a quantitative study on all the benchmark data sets. The experimental settings are summarized in Table II. The proposed criteria are compared with the LOO error rate in [11] and the fivefold CV error rate in terms of the number of function evaluations in model selection, the model selection time, and the classification error rate obtained by the KLDA using the selected model. Both cases of tuning of single and multiple kernel parameters are evaluated. Because the model selection time is affected by the optimization method, this factor has to be considered for a fair comparison. In the experiments, the comparison is conducted by using three different optimization methods: 1) the Matlab function fminsearch(), 2) the Matlab function fminunc() with $\mathrm{GradObj} =$ "off," that is, gradient information is used in optimization and it is computed by the function fminunc() itself, and 3) the Matlab function fminunc() with $\mathrm{GradObj} =$ "on," where the gradient information is computed by the user and input into fminunc() as an argument. The function fminsearch()

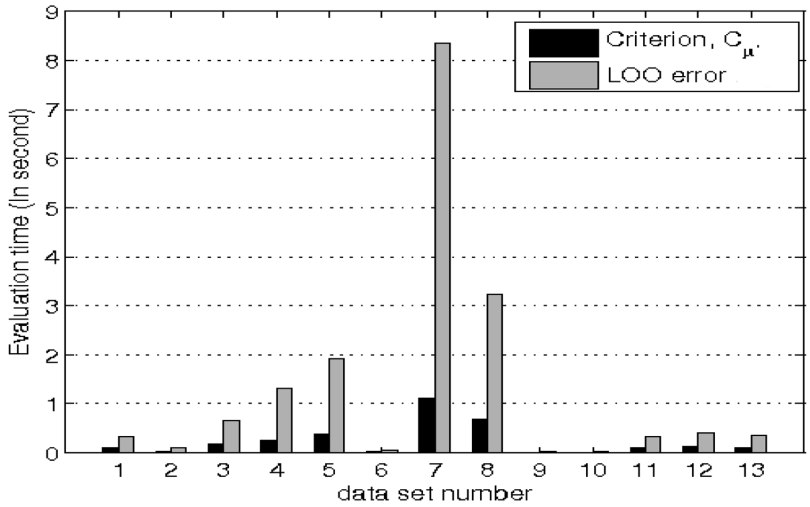


Fig. 4. Comparison of the time for a single evaluation of the proposed criterion or the LOO error rate in [11]. Each evaluation of the proposed criterion costs less time, showing its advantage on computational efficiency.

TABLE II
SUMMARY OF EXPERIMENTAL SETTINGS

Result	Involved criteria	Kernel parameter(s)	Comparison of	Optimization method
Table III	$\mathcal{C}_{\mu'}$, $\mathcal{C}_l$, LOO	single	#f_eval, time	fminsearch()
Table IV	$\mathcal{C}_{\mu'}$, $\mathcal{C}_l$, LOO, 5foldCV	single	classification error	fminsearch()
Table V	$\mathcal{C}_{\mu'}$, $\mathcal{C}_l$, LOO	single	#f_eval, time	fminunc(), GradObj = off
Table VI	$\mathcal{C}_{\mu'}$, $\mathcal{C}_l$, LOO, 5foldCV	single	classification error	fminunc(), GradObj = off
Table VII	$\mathcal{C}_{\mu'}$, $\mathcal{C}_l$, LOO	single	#f_eval, time	fminunc(), GradObj = on
Table VIII	$\mathcal{C}_{\mu'}$, $\mathcal{C}_l$, LOO, 5foldCV	single	classification error	fminunc(), GradObj = on
Table IX	$\mathcal{C}_{\mu'}$, $\mathcal{C}_l$, LOO	multiple	#f_eval, time	fminsearch()
Table X	$\mathcal{C}_{\mu'}$, $\mathcal{C}_l$, LOO, 5foldCV	multiple	classification error	fminsearch()
Table XI	$\mathcal{C}_{\mu'}$, $\mathcal{C}_l$, LOO	multiple	#f_eval, time	fminunc(), GradObj = off
Table XII	$\mathcal{C}_{\mu'}$, $\mathcal{C}_l$, LOO, 5foldCV	multiple	classification error	fminunc(), GradObj = off
Table XIII	$\mathcal{C}_{\mu'}$, $\mathcal{C}_l$, LOO	multiple	#f_eval, time	fminunc(), GradObj = on
Table XIV	$\mathcal{C}_{\mu'}$, $\mathcal{C}_l$, LOO, 5foldCV	multiple	classification error	fminunc(), GradObj = on
Table XV	$\mathcal{C}_{\mu'}$	multiple	two settings of θ_0 in (22)	fminunc(), GradObj = on

⋆ LOO: The leave-one-out cross-validation error rate; 5foldCV: five-fold cross-validation error rate
⋆ #f_eval: Number of function evaluations in model selection.

implements a Nelder–Mead simplex method that does not use gradient information. The function fminunc() implements a Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton method that makes use of gradient information. For each of the Matlab functions, its default optimization setting is used and no extra measure is taken to speed up the optimization. For each data set, model selection is individually performed on each of the predefined 100 or 20 training and test subsets, and the averaged results are used for comparison.

B. Comparison on Tuning a Single Kernel Parameter

This part includes six tables (Tables III–VIII). They form three groups, each of which corresponds to one optimization method. In each group, there are two tables. One of them compares the model selection time, and the other compares the classification error rates. The details on the experimental settings can be found from Table II.

The proposed variants of the criterion and the LOO error rate are first compared by using fminsearch() for optimization. The total number of function evaluations (denoted by #f_eval) and the model selection time are listed in Table III. For $\mathcal{C}_{\mu'}$ and $\mathcal{C}_l$, the result is the addition of two parts: 1) the time taken for tuning the kernel parameter, and 2) the time taken for tuning the regularization parameter μ by using the method in [1]. As for the LOO error rate in [11], it tunes both the kernel parameter and the regularization parameter in a single optimization. By comparing the model selection time taken by each criterion, it can be clearly seen that both $\mathcal{C}_{\mu'}$ and $\mathcal{C}_l$ produce a faster model selection than the LOO error rate in [11]. The model selection time can be reduced up to five or six times in general. Especially, for the data sets of "Image" and "Splice" that have a larger number of training samples, the reduction of model selection time is more significant in absolute terms. These results are consistent with that in Fig. 4, as well as the previous analysis that our criterion does not involve any matrix-inverse operation and thus can be computed with less computational overhead.

Now let us check whether the model selected by the proposed criterion can give rise to good classification performance. The KLDA is performed by using the model parameters selected by the proposed criterion, the fivefold CV error rate, and the LOO error rate, respectively. With the KLDA, both training and test

TABLE III
COMPARISON OF MODEL SELECTION TIME (fminsearch(), TUNING A SINGLE KERNEL PARAMETER)

Data set	$\mathcal{C}_{\mu'}$ + method [1]		$\mathcal{C}_l$ + method [1]		LOO error [11]	
	#f_eval	Time	#f_eval	Time	#f_eval	Time
Banana	37.7 + 35.9	1.5 + 2.5	38.0 + 37.0	1.5 + 2.6	105.7	23.1
B.Cancer	46.2 + 32.0	0.4 + 0.3	33.7 + 41.1	0.4 + 0.4	118.8	4.5
Diabetes	44.0 + 30.0	1.9 + 3.2	41.8 + 41.1	2.2 + 4.6	107.1	34.4
Flare Solar	36.7 + 41.1	3.7 + 12.0	38.8 + 41.8	4.2 + 12.2	117.0	95.1
German	40.0 + 31.9	3.9 + 10.8	23.4 + 41.3	2.9 + 13.9	99.5	92.6
Heart	42.0 + 36.5	0.2 + 0.2	30.0 + 40.5	0.2 + 0.3	123.9	3.1
Image	40.0 + 42.8	12.9 + 92.0	35.7 + 32.1	14.6 + 66.1	110.0	559.6
Splice	38.0 + 38.0	7.3 + 36.9	28.0 + 38.0	7.0 + 36.0	103.2	255.9
Thyroid	35.2 + 35.0	0.1 + 0.2	39.3 + 38.5	0.2 + 0.1	101.3	1.7
Titanic	39.9 + 33.5	0.2 + 0.1	25.4 + 38.4	0.1 + 0.2	146.5	7.2
Ringnorm	36.0 + 40.0	1.4 + 2.7	38.0 + 40.0	1.5 + 2.8	112.8	24.3
Twonorm	38.0 + 37.5	1.2 + 2.6	32.3 + 39.3	1.3 + 2.7	99.9	21.7
Wavenorm	38.0 + 34.6	1.2 + 2.4	32.0 + 42.0	1.3 + 2.7	106.0	23.1

⋆ LOO error: The leave-one-out cross-validation error rate.
⋆ Time unit: Second.

TABLE IV
CLASSIFICATION ERROR AFTER THE KLDA (fminsearch(), TUNING A SINGLE KERNEL PARAMETER)

Data set	$\mathcal{C}_{\mu'}$ + method [1]				$\mathcal{C}_l$ + method [1]				5-fold CV	LOO error in [11]
	Test error	McNemar ↑	↓	=	Test error	McNemar ↑	↓	=	Test error	Test error
Banana	10.6±0.5	42	1	57	10.6±0.5	41	2	57	11.1±0.7	10.7±0.6
B.Cancer	26.8±4.6	1	0	99	26.4±4.2	1	0	99	27.4±4.8	26.8±4.5
Diabetes	23.2±1.8	3	0	97	23.9±1.9	0	3	97	23.6±1.7	23.4±1.9
Flare Solar	33.9±1.6	3	1	96	34.0±1.7	4	2	94	34.3±1.7	34.1±1.5
German	23.7±2.1	1	1	98	23.1±2.1	3	0	97	23.5±2.0	23.1±2.2
Heart	16.7±3.6	3	0	97	17.1±3.7	3	0	97	17.8±3.8	16.8±3.7
Image▽	7.3±0.7	0	20	0	3.5±0.6	0	3	17	2.9±0.6	3.3±0.7
Splice	11.7±0.8	0	9	11	11.2±0.7	0	1	19	10.9±0.7	10.9±0.7
Thyroid	3.6±1.9	1	0	99	3.6±1.8	0	0	100	4.3±2.1	4.1±2.3
Titanic	22.8±1.1	10	37	53	22.7±1.0	9	29	62	22.4±1.0	22.6±1.3
Ringnorm	1.7±0.1	29	1	70	1.7±0.2	34	2	64	1.8±0.2	1.6±0.1
Twonorm	2.5±0.2	27	4	69	2.7±0.2	18	26	56	2.7±0.3	2.6±0.2
Waveform	9.7±0.4	19	4	77	9.6±0.3	27	1	72	9.9±0.6	9.6±0.4

↑ : on the indicated number of subsets, the result of the proposed criterion is statistically *better* than that of the 5-fold CV;
↓ : on the indicated number of subsets, the result of the proposed criterion is statistically *worse* than that of the 5-fold CV;
= : *No statistical difference* between the results from the proposed criterion and the 5-fold CV

TABLE V
COMPARISON OF MODEL SELECTION TIME (fminunc(), GradObj = "off," TUNING A SINGLE KERNEL PARAMETER)

Data set	$\mathcal{C}_{\mu'}$ + method [1]		$\mathcal{C}_l$ + method [1]		LOO error [11]	
	#f_eval	Time	#f_eval	Time	#f_eval	Time
Banana	18.6 + 35.9	0.9 + 2.5	17.9 + 37.0	0.9 + 2.5	70.8	16.9
B.Cancer	16.7 + 32.0	0.2 + 0.3	12.7 + 41.1	0.2 + 0.4	86.4	2.9
Diabetes	17.2 + 30.0	0.9 + 3.2	14.8 + 41.1	1.0 + 4.4	74.5	25.6
Flare Solar	17.1 + 41.1	2.1 + 12.0	14.7 + 41.8	1.9 + 12.2	98.5	91.4
German	18.1 + 31.9	2.2 + 10.8	14.6 + 41.2	2.2 + 13.9	107.6	114.4
Heart	15.1 + 36.5	0.1 + 0.2	10.7 + 40.5	0.1 + 0.3	82.0	1.9
Image	16.9 + 42.8	6.6 + 87.7	16.5 + 32.1	8.0 + 65.6	72.0	453.7
Splice	16.7 + 38.0	4.1 + 35.9	15.9 + 38.0	4.7 + 36.0	105.6	309.6
Thyroid	13.9 + 35.0	0.1 + 0.1	12.9 + 38.5	0.1 + 0.1	63.4	0.9
Titanic	17.6 + 33.6	0.1 + 0.1	13.8 + 38.8	0.1 + 0.2	85.2	1.5
Ringnorm	14.9 + 40.0	0.7 + 2.7	18.6 + 40.0	0.9 + 2.7	99.8	23.2
Twonorm	16.2 + 37.5	0.6 + 2.6	14.8 + 39.3	0.7 + 2.7	88.0	20.7
Wavenorm	16.0 + 34.6	0.6 + 2.4	13.8 + 40.0	0.7 + 2.7	73.5	16.9

⋆ LOO error: The leave-one-out cross-validation error rate.
⋆ Time unit: Second.

data are projected to a 1-D subspace. A Bayes classifier is then trained in the subspace by modeling each class as a Gaussian distribution. The average classification error rates (with the standard deviation) are compared. The classification result obtained by using the model selected by the fivefold CV is used as a benchmark. To give a quantitative measure, the McNemar test (with significance level of 0.05) [24] is used to detect whether a statistically significant difference exists between the classifi-

TABLE VI
CLASSIFICATION ERROR AFTER THE KLDA (fminunc(), $\mathrm{GradObj} =$ "off," TUNING A SINGLE KERNEL PARAMETER)

Data set	$\mathcal{C}_{\mu'}$ + method [1]	McNemar			$\mathcal{C}_l$ + method [1]	McNemar			5-fold CV	LOO error in [11]
	Test error	↑	↓	=	Test error	↑	↓	=	Test error	Test error
Banana	10.6±0.5	42	1	57	10.6±0.5	41	2	57	11.1±0.7	10.6±0.4
B.Cancer	26.8±4.6	1	0	99	26.4±4.2	1	0	99	27.4±4.8	26.7±4.6
Diabetes	23.2±1.8	3	0	97	23.9±1.9	0	3	97	23.6±1.7	23.3±1.9
Flare Solar	33.9±1.6	3	1	96	34.0±1.7	4	2	94	34.3±1.7	34.3±1.7
German	23.7±2.1	1	1	98	23.1±2.1	3	0	97	23.5±2.0	23.2±2.1
Heart	16.7±3.6	3	0	97	17.1±3.7	3	0	97	17.8±3.8	17.9±5.0
Image▽	7.3±0.7	0	20	0	3.5±0.6	0	3	17	2.9±0.6	3.1±0.4
Splice	11.7±0.8	0	9	11	11.1±0.7	0	1	19	10.9±0.7	11.3±0.8
Thyroid	3.6±1.9	1	0	99	3.6±1.8	0	0	100	4.3±2.1	4.1±2.1
Titanic	22.8±1.1	10	37	53	22.8±1.4	8	29	63	22.4±1.0	22.6±1.2
Ringnorm	1.7±0.1	29	1	70	1.7±0.2	34	2	64	1.8±0.2	1.6±0.1
Twonorm	2.5±0.2	27	4	69	2.7±0.2	18	26	56	2.7±0.3	2.6±0.2
Waveform	9.7±0.4	19	4	77	9.6±0.3	21	7	72	9.9±0.6	9.6±0.4

↑ : on the indicated number of subsets, the result of the proposed criterion is statistically *better* than that of the 5-fold CV;
↓ : on the indicated number of subsets, the result of the proposed criterion is statistically *worse* than that of the 5-fold CV;
= : *No statistical difference* between the results from the proposed criterion and the 5-fold CV

TABLE VII
COMPARISON OF MODEL SELECTION TIME (fminunc(), $\mathrm{GradObj} =$ "on" TUNING A SINGLE KERNEL PARAMETER)

Data set	$\mathcal{C}_{\mu'}$ + method [1]		$\mathcal{C}_l$ + method [1]		LOO error [11]	
	#f_eval	Time	#f_eval	Time	#f_eval	Time
Banana	9.3 + 35.9	0.6 + 2.5	8.9 + 37.0	0.6 + 2.5	23.0	68.6
B.Cancer	8.4 + 32.0	0.1 + 0.3	5.0 + 41.1	0.1 + 0.4	29.6	6.6
Diabetes	8.1 + 30.0	0.7 + 3.2	6.5 + 41.1	0.6 + 4.4	23.3	112.2
Flare Solar	8.1 + 41.1	1.5 + 12.0	5.4 + 41.8	1.1 + 12.2	24.5	330.9
German	9.0 + 31.9	1.7 + 10.7	5.0 + 41.3	1.1 + 13.9	30.9	379.6
Heart	7.5 + 36.5	0.1 + 0.2	5.0 + 40.5	0.1 + 0.2	29.4	6.9
Image	7.9 + 43.1	4.9 + 88.2	6.1 + 32.6	4.4 + 65.7	16.9	1810.1
Splice	8.1 + 38.0	3.1 + 35.9	7.0 + 38.0	3.0 + 36.0	23.4	1151.7
Thyroid	6.9 + 35.0	0.1 + 0.1	6.5 + 38.5	0.1 + 0.1	19.8	2.5
Titanic	8.4 + 33.5	0.1 + 0.1	5.4 + 38.5	0.1 + 0.2	28.3	4.9
Ringnorm	7.4 + 40.0	0.5 + 2.7	8.9 + 40.0	0.6 + 2.7	33.5	88.3
Twonorm	8.0 + 37.5	0.5 + 2.6	7.0 + 39.3	0.5 + 2.7	25.7	68.6
Wavenorm	8.0 + 34.6	0.5 + 2.4	6.4 + 40.0	0.5 + 2.7	23.5	61.9

⋆ LOO error: The leave-one-out cross-validation error rate.
⋆ Time unit: Second.

cation error rate from the proposed criterion and that from the fivefold CV. Table IV reports the comparison result. It consists of four parts, showing the classification error rates of the KLDA using the models selected by different criteria. In the first two parts, besides the classification error rates, the McNemar test result on the 100 or 20 predefined test subsets is summarized for each data set. A McNemar test result has three measures. The "↑" means that on the indicated number of test subsets, the classification result of the proposed criterion is statistically *better* than that from the fivefold CV. In other words, a statistically significant difference is detected between them and the classification error rate obtained by using the proposed criterion is lower. Similarly, the "↓" means that the classification result of the proposed criterion is statistically *worse*, and the "=" means that *no statistically significant difference* is detected. As shown in Table IV, the McNemar test result suggests that the difference between the classification results is insignificant on most data sets. On the data sets of "Banana," "Ringnorm," "Twonorm" (for $\mathcal{C}_{\mu'}$ only), and "Waveform," our criteria produce slightly better performance. Meanwhile, on "Titanic," "Twonorm" (for $\mathcal{C}_l$ only), and "Splice," their performance is slightly worse. On the data set of "Image" (marked by "▽"), the criterion $\mathcal{C}_{\mu'}$ fails to select a reasonable model and the number under "↓" dominates. The two criteria $\mathcal{C}_{\mu'}$ and $\mathcal{C}_l$ give similar classification performance on all the data sets except for "Image," "Splice," and "Twonorm." In addition, by comparing the proposed criterion with the LOO error rate in [11], it can be observed that they are comparable on most data sets but the LOO error rate is slightly better on "Image" and "Splice."

These criteria are further compared by using fminunc() with $\mathrm{GradObj} =$ "off" as the optimization method. In this method, gradient information is used in optimization and it is computed by fminunc() itself. The model selection time is compared in Table V. Our criteria still cost less model selection time than the LOO error rate. Compared with the case of using fminsearch(), optimizing the proposed criterion with fminunc() requires less function evaluations and takes less time. The classification error rates are compared in Table VI. They are almost the same as those obtained in Table IV where fminsearch() is used. This suggests that both fminsearch() and fminunc() (with $\mathrm{GradObj} =$ "off") can be used to optimize the model selection criteria and

TABLE VIII
CLASSIFICATION ERROR AFTER THE KLDA (fminunc(), GradObj = "on," TUNING A SINGLE KERNEL PARAMETER)

	$\mathcal{C}_{\mu'}$ + method [1]				$\mathcal{C}_l$ + method [1]				5-fold CV	LOO error
		McNemar				McNemar				in [11]
Data set	Test error	↑	↓	=	Test error	↑	↓	=	Test error	Test error
Banana	10.6±0.5	42	1	57	10.6±0.5	41	2	57	11.1±0.7	10.6±0.5
B.Cancer	26.8±4.6	1	0	99	26.4±4.2	1	0	99	27.4±4.8	26.6±4.6
Diabetes	23.2±1.8	3	0	97	23.9±1.9	0	3	97	23.6±1.7	23.4±1.9
Flare Solar	33.9±1.6	3	1	96	34.0±1.7	4	2	94	34.3±1.7	34.2±1.5
German	23.7±2.1	1	1	98	23.1±2.1	3	0	97	23.5±2.0	23.1±2.2
Heart	16.7±3.6	3	0	97	17.1±3.7	3	0	97	17.8±3.8	17.5±5.4
Image▽	7.3±0.7	0	20	0	3.5±0.6	0	3	17	2.9±0.6	3.3±0.8
Splice	11.7±0.8	0	9	11	11.2±0.7	0	1	19	10.9±0.7	11.0±0.8
Thyroid	3.6±1.9	1	0	99	3.6±1.8	0	0	100	4.3±2.1	4.0±2.1
Titanic	22.8±1.1	10	37	53	22.8±1.4	8	29	63	22.4±1.0	22.6±1.1
Ringnorm	1.7±0.1	29	1	70	1.7±0.2	34	2	64	1.8±0.2	1.6±0.1
Twonorm	2.5±0.2	27	4	69	2.7±0.2	18	26	56	2.7±0.3	2.6±0.2
Waveform	9.7±0.4	19	4	77	9.6±0.3	27	1	72	9.9±0.6	9.6±0.4

↑ : on the indicated number of subsets, the result of the proposed criterion is statistically *better* than that of the 5-fold CV;
↓ : on the indicated number of subsets, the result of the proposed criterion is statistically *worse* than that of the 5-fold CV;
= : *No statistical difference* between the results from the proposed criterion and the 5-fold CV

TABLE IX
COMPARISON OF MODEL SELECTION TIME (fminsearch(), TUNING MULTIPLE KERNEL PARAMETERS)

	$\mathcal{C}_{\mu'}$ + method [1]		$\mathcal{C}_l$ + method [1]		LOO error [11]	
Data set	#f_eval	Time	#f_eval	Time	#f_eval	Time
Banana	88.7 + 35.9	3.7 + 2.6	91.9 + 37.0	3.7 + 2.7	199.6	32.5
B.Cancer	395.1 + 32.1	6.0 + 0.3	274.7 + 41.1	4.8 + 0.4	1865.7	118.4
Diabetes	298.2 + 30.0	33.9 + 3.4	278.4 + 41.1	34.3 + 4.6	1633.2	515.8
Flare Solar	301.0 + 41.1	81.1 + 12.5	284.8 + 41.8	77.5 + 12.9	1598.7	2873.0
German	1214.3 + 32.0	761.8 + 11.3	792.0 + 41.3	532.0 + 14.6	4200.2	5367.7
Heart	922.8 + 36.9	13.8 + 0.2	492.8 + 40.5	7.2 + 0.3	2800.2	126.2
Image	900.7 + 43.4	1915.4 + 140.7	611.7 + 32.1	1351.3 + 103.8	3801.0	25648.0
Splice†	304.9 + 38.0	1290.1 + 18.2	292.1 + 38.0	1305.5 + 18.2	265.1	1990.0
Thyroid	182.6 + 35.0	1.1 + 0.1	198.8 + 38.5	1.3 + 0.2	1001.5	18.2
Titanic	305.6 + 39.6	1.4 + 0.2	97.4 + 38.3	0.6 + 0.2	439.8	13.4
Ringnorm	668.8 + 40.0	148.7 + 2.9	700.4 + 40.0	165.4 + 2.9	4200.3	1521.1
Townorm	852.1 + 37.5	193.3 + 2.7	524.3 + 39.3	115.3 + 2.8	4200.2	1813.8
Wavenorm	1145.8 + 34.5	271.4 + 2.5	657.5 + 40.0	161.9 + 2.8	4400.2	1757.9

⋆ LOO error: The leave-one-out cross-validation error rate.
⋆ Time unit: Second.
† On "Splice": the fminsearch() is initialized with the model parameters selected by the 5-fold CV and the options.MaxIter is limited to 200. This is equally applied to the proposed criteria and the LOO error rate.

they lead to similar classification performance. However, using fminunc() can achieve a faster model selection.

Finally, these model selection criteria are compared again by using fminunc() with GradObj = "on." In this optimization method, the gradient information is computed by the user and then input into fminunc() as an argument. The model selection time is reported in Table VII. Our criteria and the LOO error rate respond to the change of the setting of GradObj differently. For both $\mathcal{C}_{\mu'}$ and $\mathcal{C}_l$, the number of function evaluations and the model selection time drop further when compared with the case of GradObj = "off." The reduction of the number of function evaluations is due to the fminunc() not calculating the gradient information by itself anymore and this saves many function evaluations. The reduction of the model selection time indicates that for $\mathcal{C}_{\mu'}$ and $\mathcal{C}_l$, analytically computing its gradient information by the user is computationally more efficient than letting fminunc() compute this by itself (for example, via finite difference). As for the LOO error rate, its model selection time significantly increases although the number of function evaluations decreases. This is because the computation of the LOO error rate is more complicated than that of the proposed criterion. Each evaluation of its gradient information requires a number of matrix operations and this prolongs the model selection process. The comparison of classification error rates is presented in Table VIII. The results are the same as those obtained in the previous experiments.

C. Comparison on Tuning Multiple Kernel Parameters[3]

In this part, the *ellipsoidal* GRBF kernel is used. It assigns each feature component an individual kernel parameter. The multiple kernel parameters are tuned in model selection. As before, the model selection time and the classification error rate are compared by using three different optimization methods.

Tables IX and X report the result when fminsearch() is used as the optimization method. Compared with its counterpart for tuning a single kernel parameter (in Table III), the number of

[3]Please note that in the case of tuning multiple kernel parameters, some model selection results for "Image" and "Splice" are obtained from part of the 20 training and test subsets. This is because the model selection time on the two data sets are relatively long and we only test (not selectively) part of the subsets.

TABLE X
CLASSIFICATION ERROR AFTER THE KLDA (fminsearch(), TUNING MULTIPLE KERNEL PARAMETERS)

	$\mathcal{C}_{\mu'}$ + method [1]				$\mathcal{C}_l$ + method [1]				5-fold CV	LOO error in [11]
		McNemar				McNemar				
Data set	Test error	↑	↓	=	Test error	↑	↓	=	Test error	Test error
Banana	10.6±0.5	42	2	56	10.6±0.5	41	2	57	11.1±0.8	10.7±0.6
B.Cancer	26.8±4.7	1	0	99	26.4±4.2	1	0	99	27.4±4.7	29.9±1.3
Diabetes	23.2±1.8	3	0	97	23.9±1.9	0	3	97	23.6±1.7	23.9±2.0
Flare Solar	33.9±1.6	3	1	96	34.0±1.7	3	2	95	34.2±1.7	34.2±1.5
German	23.7±2.1	1	1	98	23.1±2.1	1	0	99	23.5±2.0	25.0±2.4
Heart	16.6±3.6	3	0	97	17.1±3.7	3	0	97	17.8±3.8	20.3±5.2
Image▽	7.3±0.7	0	20	0	3.5±0.6	0	3	17	2.9±0.6	2.1±0.5
Splice	11.0±0.7	0	0	20	10.8±0.7	0	0	20	10.9±0.7	10.7±0.7
Thyroid	3.6±1.9	1	0	99	3.6±1.8	0	0	100	4.3±2.1	4.3±2.2
Titanic	22.5±0.5	11	29	60	22.7±1.0	8	29	63	22.4±1.0	22.9±1.7
Ringnorm	1.7±0.1	29	0	71	1.7±0.2	37	2	61	1.8±0.2	3.4±1.1
Twonorm	2.5±0.2	27	5	68	2.7±0.2	18	26	56	2.7±0.3	4.2±0.9
Waveform	9.7±0.4	18	6	76	9.6±0.3	30	0	70	9.9±0.6	11.6±1.1

↑ : on the indicated number of subsets, the result of the proposed criterion is statistically *better* than that of the 5-fold CV;
↓ : on the indicated number of subsets, the result of the proposed criterion is statistically *worse* than that of the 5-fold CV;
= : *No statistical difference* between the results from the proposed criterion and the 5-fold CV

TABLE XI
COMPARISON OF MODEL SELECTION TIME (fminunc(), $\mathrm{GradObj} =$ "off," TUNING MULTIPLE KERNEL PARAMETERS)

	$\mathcal{C}_{\mu'}$ + method [1]		$\mathcal{C}_l$ + method [1]		LOO error [11]	
Data set	#f_eval	Time	#f_eval	Time	#f_eval	Time
Banana	29.1 + 35.9	1.9 + 3.0	35.9 + 37.0	2.6 + 3.1	133.1	32.0
B.Cancer	87.1 + 32.1	2.8 + 0.4	52.7 + 41.1	2.6 + 0.5	852.1	136.3
Diabetes	57.8 + 30.0	13.8 + 4.0	60.2 + 41.1	16.6 + 5.4	709.4	296.1
Flare Solar	47.3 + 41.1	32.7 + 17.5	66.6 + 41.8	40.6 + 12.2	801.9	4004.7
German	192.8 + 32.0	304.7 + 15.6	141.2 + 41.3	269.8 + 14.2	2112.0	2702.4
Heart	160.6 + 36.9	6.3 + 0.3	83.6 + 40.5	5.2 + 0.2	1288.1	59.2
Image	147.8 + 43.1	705.0 + 137.9	110.5 + 32.1	704.8 + 103.7	1900.0	15402.2
Splice	1210.3 + 39.6	15185.4 + 25.0	860.2 + 38.0	13053.8 + 36.3	3720.0	45821.8
Thyroid	37.8 + 35.0	0.8 + 0.2	47.7 + 38.5	1.0 + 0.1	421.6	7.6
Titanic	37.8 + 33.6	0.6 + 0.2	31.2 + 39.0	0.7 + 0.2	238.8	9.0
Ringnorm	77.9 + 40.0	71.7 + 3.4	105.5 + 40.0	77.1 + 2.8	1690.1	630.2
Twonorm	122.9 + 37.5	76.5 + 3.1	98.7 + 39.3	76.0 + 2.8	1672.0	582.9
Wavenorm	151.7 + 34.5	92.0 + 2.9	101.5 + 40.0	86.0 + 2.7	2197.0	1243.2

⋆ LOO error: The leave-one-out cross-validation error rate.
⋆ Time unit: Second.

function evaluations and the model selection time increase drastically. This is not surprising because the number of kernel parameters to be optimized is much larger than before. By comparing the proposed criterion with the LOO error rate, it can still be seen that the former costs much less mode selection time, especially on the data sets of "Ringnorm," "Twonorm," "Waveform," and "Image" in which the data sets have a large number of training samples or a high-dimensional input space. As shown in Table X, the significance test result indicates that our criteria still work well for tuning multiple kernel parameters. More importantly, similar classification performance can be achieved by merely using a part of features automatically selected by tuning multiple kernel parameters. For instance, on "Breast Cancer," only two out of nine features are assigned with nonzeros η_i [recall that $\eta_i = 1/(2\sigma_i^2)$]. On "Titanic," by using $\mathcal{C}_{\mu'}$, only the third feature (it indicates the gender of a passenger on "Titanic") is assigned nonzero η_i. With this model, the KLDA achieves a lower error rate. This suggests that multiparameter-based model selection can possibly be used to identify important features before applying KLDA. The LOO error rate still demonstrates good performance except on "Heart," "Ringnorm," "Twonorm," and "Waveform." However, it may be too premature to conclude that the LOO error rate cannot work well for the case of multiple kernel parameter tuning. The work in [11] focuses on tuning a single kernel parameter, and the LOO error rate in that work has not incorporated the regularization term that is often needed in a multiparameter optimization problem. It could be expected that better performance may be attained when suitable regularization is imposed. However, this is beyond the scope of this paper.

The result of employing fminunc() with $\mathrm{GradObj} =$ "off" for optimization is presented in Tables XI and XII. By replacing fminsearch() with fminunc(), our criteria need less model selection time. Compared with the LOO error rate, they still achieve a faster model selection. The classification performance given by the selected models is still comparable to that of the fivefold CV (except for the data set of "Image"). By setting $\mathrm{GradObj} =$ "on," our criteria are compared with the LOO error rate and the fivefold CV again in Tables XIII and XIV. The model selection time taken by the proposed criterion is further reduced, whereas the time taken by the LOO error rate significantly increases. As explained earlier, this is because analytically computing the gradient information of the LOO error rate is computationally expensive.

TABLE XII
CLASSIFICATION ERROR AFTER THE KLDA (fminunc(), GradObj = "off," TUNING MULTIPLE KERNEL PARAMETERS)

	$\mathcal{C}_{\mu'}$ + method [1]				$\mathcal{C}_l$ + method [1]				5-fold CV	LOO error
		McNemar				McNemar				in [11]
Data set	Test error	↑	↓	=	Test error	↑	↓	=	Test error	Test error
Banana	10.6±0.5	42	2	56	10.6±0.5	41	2	57	11.1±0.8	10.7±0.6
B.Cancer	26.8±4.7	1	0	99	26.7±3.9	1	0	99	27.4±4.7	28.4±3.4
Diabetes	23.2±1.8	3	0	97	23.7±1.8	0	3	97	23.6±1.7	22.5±1.5
Flare Solar	33.8±1.6	3	1	96	34.4±1.8	3	2	95	34.2±1.7	34.9±1.9
German	23.7±2.1	1	1	98	22.4±2.5	1	0	99	23.5±2.0	24.6±2.0
Heart	16.6±3.6	3	0	97	16.6±3.6	3	0	97	17.8±3.8	19.6±3.5
Image▽	7.3±0.7	0	20	0	3.8±0.5	0	3	17	2.9±0.6	1.9±0.3
Splice▽	13.2±6.2	9	4	3	9.8±0.8	4	0	2	10.9±0.7	9.1±0.0
Thyroid	3.6±1.9	1	0	99	3.5±1.7	0	0	100	4.3±2.1	4.3±2.1
Titanic	22.6±0.5	11	29	60	22.6±1.2	9	29	62	22.4±1.0	21.9±1.0
Ringnorm	1.7±0.1	29	0	71	1.7±0.2	37	2	61	1.8±0.2	2.0±0.4
Twonorm	2.5±0.2	27	5	68	2.7±0.2	18	26	56	2.7±0.3	4.9±0.7
Waveform	9.7±0.4	18	6	76	9.6±0.3	30	0	70	9.9±0.6	12.1±1.3

↑ : on the indicated number of subsets, the result of the proposed criterion is statistically *better* than that of the 5-fold CV;
↓ : on the indicated number of subsets, the result of the proposed criterion is statistically *worse* than that of the 5-fold CV;
= : *No statistical difference* between the results from the proposed criterion and the 5-fold CV

TABLE XIII
COMPARISON OF MODEL SELECTION TIME (fminunc(), GradObj = "on," TUNING MULTIPLE KERNEL PARAMETERS)

	$\mathcal{C}_{\mu'}$ + method [1]		$\mathcal{C}_l$ + method [1]		LOO error [11]	
Data set	#f_eval	Time	#f_eval	Time	#f_eval	Time
Banana	14.7 + 35.9	1.2 + 2.9	14.9 + 37.0	1.3 + 1.4	33.4	154.5
B.Cancer	15.4 + 32.1	0.9 + 0.4	9.0 + 40.9	0.8 + 0.3	88.5	255.9
Diabetes	12.7 + 29.7	4.9 + 3.7	11.5 + 41.1	5.3 + 2.5	77.0	2054.3
Flare Solar	11.6 + 41.1	10.9 + 13.7	10.2 + 41.8	12.1 + 6.5	85.9	10070.8
German	17.2 + 32.0	55.7 + 12.3	11.0 + 41.3	50.4 + 7.2	166.8	34455.8
Heart	17.5 + 36.8	1.1 + 0.3	10.2 + 39.8	0.9 + 0.2	98.0	242.0
Image	37.7 + 43.0	262.0 + 57.9	37.2 + 31.9	279.5 + 68.4	102.0	114051.9
Splice	27.1 + 39.2	1775.2 + 18.1	20.3 + 38.0	1654.8 + 17.5	-	-
Thyroid	10.9 + 35.0	0.2 + 0.2	12.0 + 38.4	0.2 + 0.1	62.2	26.2
Titanic	34.4 + 41.2	0.3 + 0.2	9.6+ 38.8	0.2 + 0.2	45.3	14.8
Ringnorm	10.2 + 40.0	14.5 + 3.2	12.8 + 40.0	15.3 + 1.5	90.6	3574.7
Twonorm	13.2 + 37.5	15.5 + 3.0	11.0 + 38.8	15.1 + 1.7	80.2	4905.4
Wavenorm	14.2 + 34.5	17.7 + 2.8	10.4 + 40.0	16.9 + 2.0	121.0	5464.1

⋆ LOO error: The leave-one-out cross-validation error rate.
⋆ Time unit: Second.

D. Summary of the Experimental Results

For both single and multiple kernel parameter tuning, our criteria consistently achieve a faster model selection when different optimization methods are used. In terms of model selection time, our criteria work best with the optimization method of fminunc() with GradObj = "on," whereas the LOO error rate works best with fminunc() with GradObj = "off." For the classification error rate, the McNemar test confirms that on six out of 13 data sets (Breast Cancer, Diabetes, Flare Solar, German, Heart, and Thyroid), there is no significant difference between the classification error rates obtained with the model selected by the proposed criterion and that from the fivefold CV. By comparing the classification results obtained by using $\mathcal{C}_{\mu'}$ and $\mathcal{C}_l$, it can be seen that they are comparable on most data sets. The criterion $\mathcal{C}_l$ may be a better choice for practical use because it does not need to empirically set an extra parameter μ'.

Before ending this section, two settings of $\boldsymbol{\theta}_0$ in (22) are compared. Setting I (proposed in this paper) applies the constraint of $\theta_1 = \cdots = \theta_d$ and solves the optimization problem in (23) to estimate $\boldsymbol{\theta}_0$. Setting II simply sets $\boldsymbol{\theta}_0 = \mathbf{0}$. With different values of the regularization parameter λ, the effects due to the two settings are compared in terms of the obtained classification error rates in Table XV. For ease of comparison, the lowest error rate (with respect to the value of λ) from each setting is highlighted in bold. As shown, Setting I achieves lower classification error rates on all the data sets except for the "Titanic" where the two settings give similar results. As mentioned above, the better performance obtained by using Setting I is because this setting first secures a good initialization and then seeks further improvement. The explanation has been given in Section III-E.

V. CONCLUSION

This paper proposes a kernel-induced space selection approach to tackle model selection in KLDA. The optimal model is regarded as the one giving rise to a feature space in which the separability of different classes is maximized. A scatter-matrix-based criterion is developed to measure the class separability in a feature space, and the optimal kernel parameters are obtained by maximizing this criterion. The computational complexity of the proposed criterion and its relationship to the KLDA are analyzed. Experimental study is conducted on a set of benchmark data sets to verify the

TABLE XIV
CLASSIFICATION ERROR AFTER THE KLDA (fminunc(), $\mathrm{GradObj}$ = "on," TUNING MULTIPLE KERNEL PARAMETERS)

	$\mathcal{C}_{\mu'}$ + method [1]				$\mathcal{C}_l$ + method [1]				5-fold CV	LOO error in [11]
		McNemar				McNemar				
Data set	Test error	↑	↓	=	Test error	↑	↓	=	Test error	Test error
Banana	10.6±0.5	42	2	56	10.6±0.5	41	2	57	11.1±0.8	10.7±0.6
B.Cancer	26.8±4.7	1	0	99	26.4±4.3	1	0	99	27.4±4.7	29.7±4.2
Diabetes	23.1±1.8	3	0	97	23.9±1.9	0	3	97	23.6±1.7	23.4±1.9
Flare Solar	33.9±1.6	3	1	96	33.9±1.6	3	2	95	34.2±1.7	34.6±1.9
German	23.7±2.1	1	1	98	23.1±2.1	1	0	99	23.5±2.0	24.5±1.8
Heart	16.6±3.6	3	0	97	17.0±4.1	3	0	97	17.8±3.8	22.3±3.7
Image▽	7.3±0.7	0	20	0	3.4±0.5	0	3	17	2.9±0.6	2.3±0.2
Splice▽	12.5±6.0	11	4	5	9.7±0.6	16	0	4	10.9±0.7	-
Thyroid	3.6±1.9	1	0	99	3.4±1.9	0	0	100	4.3±2.1	4.3±2.1
Titanic	22.6±0.5	11	29	60	22.7±1.0	8	30	62	22.4±1.0	23.0±1.3
Ringnorm	1.7±0.1	29	0	71	1.5±0.2	37	2	61	1.8±0.2	2.1±0.6
Twonorm	2.5±0.2	27	5	68	2.6±0.2	18	26	56	2.7±0.3	5.0±0.7
Waveform	9.7±0.4	18	6	76	9.5±0.3	30	0	70	9.9±0.6	12.2±1.3

↑ : on the indicated number of subsets, the result of the proposed criterion is statistically *better* than that of the 5-fold CV;
↓ : on the indicated number of subsets, the result of the proposed criterion is statistically *worse* than that of the 5-fold CV;
= : *No statistical difference* between the results from the proposed criterion and the 5-fold CV

TABLE XV
COMPARISON OF THE CLASSIFICATION ERROR RATES FROM TWO SETTINGS OF $\boldsymbol{\theta}_0$ IN (22)

	Setting I				Setting II			
Data set	$\lambda = 0.99$	0.9	0.75	0.5	$\lambda = 0.99$	0.9	0.75	0.5
Banana	**10.6±0.5**	10.6±0.5	10.6±0.5	10.6±0.5	44.6±5.3	40.1±5.9	**39.3±6.0**	39.5±6.9
B.Cancer	**26.8±4.7**	28.4±4.7	28.6±4.7	28.7±4.1	28.7±5.1	28.8±4.7	**28.7±4.2**	28.8±4.2
Diabetes	23.2±1.8	**23.1±1.8**	23.1±1.8	23.9±2.4	**25.7±2.0**	26.4±1.8	26.1±1.8	26.1±1.8
Flare Solar	**33.9±1.6**	33.9±1.6	33.9±1.6	33.9±1.6	**44.7±1.8**	44.7±1.8	44.7±1.8	44.7±1.8
German	**23.6±2.1**	30.6±2.1	30.7±2.1	30.7±2.1	**30.7±2.1**	30.7±2.1	30.7±2.1	30.7±2.1
Heart	**16.6±3.6**	24.2±4.2	24.1±3.9	24.0±4.0	**23.3±4.9**	24.2±4.2	24.0±4.0	24.1±3.9
Image	**7.3±0.7**	14.7±13.0	53.0±8.7	44.8±13.6	41.4±4.3	47.2±12.1	40.5±13.0	**34.6±9.6**
Splice	**12.5±6.0**	22.8±0.5	22.8±0.5	22.8±0.5	22.8±0.4	22.8±0.4	22.8±0.4	**22.7±0.4**
Thyroid	**3.6±1.9**	3.6±1.9	3.6±1.9	3.6±1.9	**20.0±9.3**	26.9±19.5	24.7±16.2	23.3±14.3
Titanic	22.8±1.1	22.7±1.0	22.8±0.8	**22.6±0.5**	22.8±1.0	22.9±2.4	22.8±2.4	**22.5±0.4**
Ringnorm	**1.7±0.1**	1.7±0.1	1.7±0.1	1.7±0.1	**21.1±13.1**	42.8±9.0	43.4±9.0	43.8±8.9
Twonorm	**2.5±0.2**	2.5±0.2	2.6±0.2	15.9±9.1	**3.3±0.8**	13.6±5.5	18.5±6.7	21.3±6.6
Waveform	**9.7±0.4**	10.6±1.2	16.3±1.8	19.2±2.6	**16.9±1.4**	17.5±1.7	18.6±2.0	19.8±2.3

effectiveness of the proposed approach. The following conclusions can be drawn. First, compared with the state-of-the-art method, the proposed criterion has less computational overhead and facilitates a faster model selection. When multiple kernel parameters are to be tuned or when there is a large number of training samples, the reduction of model selection time is particularly significant. Second, the model selection approach proposed in this paper can efficiently tune single and multiple kernel parameters for the KLDA. Third, an essential connection is revealed between the proposed criterion and the KLDA. It is proven to be the lower bound of the maximum value of the generalized Rayleigh quotient in KLDA's objective function. This justifies its application to model selection for KLDA and also is the reason why it works. Finally, the proposed criterion is independent of the regularization parameter in the KLDA, and hence it cannot be used to tune this parameter. This work circumvents this problem by incorporating the method in [1]. As shown in the experimental study, the regularization parameter can be efficiently optimized as soon as the kernel parameters are tuned by the proposed criterion.

The following issues are worthy of exploring in future work. It has been found that the optimized kernel parameters can reveal the importance of the features in discriminating different classes [20]. An instant application of this property is in the area of *feature selection* (a comprehensive overview can be found in [25]), that is,find $p(p < d)$ most discriminative features from the original d features while maximally maintaining the separability of classes. Some related work, such as feature scaling for the KLDA, has been developed in [12]. We think that our approach may have the advantage of computational efficiency, which allows more sophisticated feature selection strategies to be used. It is worth noting that a thorough study of feature selection with the kernel-based class separability criterion has been reported in our recent work [26]. Also, our criterion can be readily extended to multiclass classification, although this work focuses on binary classification only. In addition, the proposed criterion can be combined with the LOO error rate in [11]. When searching for the model parameter set that minimizes the LOO error rate, our criterion can be optimized first to obtain a good initialization. This may significantly shorten the model selection process using the LOO error rate while maintaining its good selection performance.

ACKNOWLEDGMENT

The authors would like to thank D. Chik for proofreading this paper.

REFERENCES

[1] K. Saadi, N. L. C. Talbot, and G. C. Cawley, "Optimally regularised kernel fisher discriminant analysis," in *Proc. 17th Int. Conf. Pattern Recognit.*, Cambridge, U.K., 2004, pp. 427–430.

[2] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Miiller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX*, Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds. Piscataway, NJ: IEEE, 1999, pp. 41–48.

[3] F. A. G. Baudat, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, no. 10, pp. 2385–2404, 2000.

[4] V. Roth and V. Steinhage, "Nonlinear discriminant analysis using kernel functions," in *Advances in Neural Information Processing Systems 568–574*. Cambridge, MA: MIT Press, 2000, vol. 12.

[5] M. Zhu and A. Martinez, "Pruning noisy bases in discriminant analysis," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 148–157, Jan. 2008.

[6] B. Schölkopf, S. Mika, C. J. C. Surges, P. Knirsch, K.-R. Muller, G. Rätsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1000–1017, Sep. 1999.

[7] J. Lu, K. Plataniotis, and A. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 117–126, Jan. 2003.

[8] Q. Liu, X. Tang, H. Lu, and S. Ma, "Face recognition using kernel scatter-difference-based discriminant analysis," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 1081–1085, Jul. 2006.

[9] S. Yang, S. Yan, C. Zhang, and X. Tang, "Bilinear analysis for kernel selection and nonlinear feature extraction," *IEEE Trans. Neural Netw.*, vol. 18, no. 5, pp. 1442–1452, Sep. 2007.

[10] S. Mika, "Kernel fisher discriminants," Ph.D. dissertation, Fakultät IV—Elektrotechnik und Informatik, Technische Univeristät, Berlin, Germany, 2002.

[11] G. C. Cawley and N. L. C. Talbot, "Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers," *Pattern Recognit.*, vol. 36, no. 11, pp. 2585–2592, 2003.

[12] L. Bo, L. Wang, and L. Jiao, "Feature scaling for kernel fisher discriminant analysis using leave-one-out cross validation," *Neural Comput.*, vol. 18, no. 4, pp. 961–978, 2006.

[13] T. P. Centeno and N. D. Lawrence, "Optimising kernel parameters and regularization coefficients for non-linear discriminant analysis," *J. Mach. Learn. Res.*, vol. 7, pp. 455–491, 2006.

[14] R. Linsker, "Self-organization in a perceptual network," *IEEE Computer*, vol. 21, no. 3, pp. 105–117, Mar. 1988.

[15] L. Wang and K. L. Chan, "Learning kernel parameters by using class separability measure," presented at the 6th Annu. Workshop Kernel Mach., Whistler, Canada, Dec. 2002.

[16] H. Xiong, M. Swamy, and M. Ahmad, "Optimizing the kernel in the empirical feature space," *IEEE Trans. Neural Netw.*, vol. 16, no. 2, pp. 460–474, Mar. 2005.

[17] R. O. Duda, D. G. Stork, and P. E. Hart, *Pattern Classification*, 2nd ed. New York: Wiley, 2001, p. 115.

[18] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. New York: Academic, 1999.

[19] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1990.

[20] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 131–159, 2002.

[21] M. G. Genton, "Classes of kernels for machine learning: A statistic perspective," *J. Mach. Learn. Res.*, no. 2, pp. 299–312, 2001.

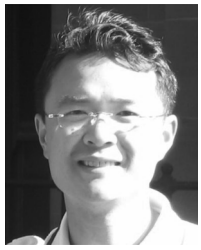
[22] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Base Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[23] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[24] T. G. Dietterich, "Approximate statistical test for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, 1998.

[25] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, pp. 131–156, 1997.

[26] L. Wang, "Feature selection with kernel class separability," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1534–1546, Sep. 2008.



Lei Wang (M'07) received the B.Eng and M.Eng degrees from the Department of Instrument Science and Engineering, Southeast University, Nanjing, China in 1996 and 1999, respectively, and the Ph.D. from School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2004.

He was a Research Associate and Research Fellow at Nanyang Technological University from 2003 to 2005. After that, he joined the Department of Information Engineering, Research School of Information Sciences and Engineering, The Australian National University, Canberra, A.C.T., Australia, as Research Fellow. In 2007, he was awarded the Australian Postdoctoral Fellowship by the Australian Research Council. His research interests include computer vision, pattern recognition, information retrieval, and machine learning.

Kap Luk Chan (S'88–M'90) received the Ph.D. degree in robot vision from Imperial College of Science, Technology and Medicine, University of London, London, U.K., in 1991.

Currently, he is an Associate Professor in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests are in image analysis and computer vision, particularly, in texture analysis, statistical image analysis, perceptual grouping, image and video retrieval, application of machine learning in computer vision, computer vision for human computer interaction, and biomedical signal and image analysis.

Dr. Chan is a member of The Institution of Engineering and Technology (IET).

Ping Xue (M'91–SM'03) received the B.S degree in electronic engineering from the Harbin Institute of Technology, Harbin, China, in 1968 and the M.S.E., M.A., and Ph.D. degrees in electrical engineering and computer science from Princeton University, Princeton, NJ, in 1981, 1982, and 1985, respectively.

He was a Member of Technical Staff of the David Sarnoff Research Center, USA, from 1984 to 1986, and on the faculty of Shanghai Jiao Tong University, China, from 1986 to 1990. He was with the Chartered Semiconductor, Singapore, from 1991 to 1994, and the Institute of Microelectronics from 1994 to 2001. He joined the Nanyang Technological University, Singapore, in 2001, as an Associate Professor. His research interests include the multimedia signal processing, content/perceptual-based analysis for video indexing and retrieval, and applications in communication networks.

Luping Zhou (M'07) received the B.Eng. degree from the Department of Instrument Science and Engineering, Southeast University, Nanjing, China, in 1996 and the M.Sc. degree from the Department of Computer Science, National University of Singapore, Singapore, 2002. Currently, she is working towards the Ph.D. degree in the Department of Information Engineering, Research School of Information Sciences and Engineering, The Australian National University, Canberra, A.C.T., Australia.

Her research interests include computer vision and machine learning, particularly, in medical image analysis.